Published by CCDCOE

2024

16th International Conference on Cyber Conflict: Over the Horizon

C. Kwan, L. Lindström,
D. Giovannelli, K. Podiņš,
D. Štrucl (Eds.)

CYCON

IEEE
Advancing Technology for Humanity

CYCON 2024: OVER THE HORIZON
16th INTERNATIONAL CONFERENCE ON CYBER CONFLICT

IEEE Catalog Number: CFP2426N-PRT
IEEE Xplore Publication: CFP2426N-ART
ISBN (print): 978-9916-9789-4-8
ISBN (pdf): 978-9916-9789-5-5

COPYRIGHT AND REPRINT PERMISSIONS

CCDCOE Publications
Filtri tee 5, 10132 Tallinn, Estonia
Phone: +372 717 6800
Fax: +372 717 6308
Email: publications@ccdcoe.org
Web: www.ccdcoe.org
Layout: JDF

NATO COOPERATIVE CYBER DEFENCE CENTRE OF EXCELLENCE

The NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) is a NATO-accredited hub of knowledge that takes a distinctive interdisciplinary approach to the most relevant issues in cyberspace. At its core, the CCDCOE is staffed by a diverse array of international experts from military, governmental, academic, and industrial backgrounds, representing 39 like-minded nations from across the globe.

The CCDCOE stands as a leading authority on cyber defence-related matters, offering invaluable expertise in strategic, legal, operational, and technical realms. It serves as a beacon of innovative thought in all facets of cyber defence, providing a comprehensive perspective on the field. The CCDCOE encourages and supports the process of mainstreaming cyber defence into NATO and national governance and capability, within its closely connected focus areas of technology, strategy, operations, and law.

The *Tallinn Manual*, developed at the initiative of CCDCOE, serves as the most comprehensive guide for legal experts and policy advisers on the application of international law to cyber operations involving states and non-state entities. The CCDCOE also co-produces the Cyber Law Toolkit – an interactive tool for all those working at or interested in the intersection of cyber operations and public international law.

Since 2010, the CCDCOE has organized the Locked Shields exercise, recognized as the largest and most complex live-fire technical cyber resilience exercise in the world. Every year, cybersecurity professionals take part in Locked Shields to hone their abilities to safeguard national IT systems and critical infrastructure amid real-time cyberattacks. The exercise is designed around realistic scenarios, cutting-edge technologies, and the simulation of a comprehensive cyber incident, encompassing strategic decision-making, legal considerations, and communication strategies.

The CCDCOE hosts the International Conference on Cyber Conflict, known as CyCon. This unique event has taken place every May in Tallinn since 2009. CyCon brings together prominent experts and decision-makers from the global cyber defence community, attracting over 600 attendees from governments, industry, and academia.

The 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.

NATO-accredited centres of excellence are not part of the NATO Command Structure.

CYCON 2024 SPONSORS

DIAMOND SPONSORS

FORTINET

Microsoft

paloalto NETWORKS

SIEMENS

Telsy | A TIM ENTERPRISE BRAND

TREND MICRO

GOLD SPONSORS

ELT GROUP

Silobreaker

TECHNICAL SPONSOR

IEEE
Advancing Technology for Humanity

TABLE OF CONTENTS

INTRODUCTION

As we move past the COVID-19 pandemic, the world is no longer the same. While we could not have expected the world of 2024 to be the same as that of 2019 and early 2020, there have been some pivotal events. The Russia–Ukraine war continues, and there is fresh conflict in the Middle East. These events have led to a paradigm shift in the nature and trajectory of cyber conflict and its underlying considerations.

Even as technological developments come at an ever-increasing pace, we are reaching a point where socio-economic and geopolitical change are also accelerating. Combined, this has the potential to cause sudden paradigm shifts in world order, participative democracy, and the structure and values of our societies and economies. This is why the Programme Committee chose 'Over the Horizon' as the theme for CyCon 2024. It denotes the need not only to look at the forthcoming technological trends on the horizon but also to start forecasting beyond that horizon due to accelerating rates of change, and to examine the outcome of the potential interplay of science, technology, innovation, industry, law, socio-economic factors, and geopolitics.

Even if predictions prove to be not completely accurate, such forecasting will help us prepare our societies and technologies for mitigating future challenges at an earlier structural stage and to maximize the exploitation of opportunities. This edition of CyCon received over 200 abstracts relating to novel research and innovation in law, technology, operations, strategy, and policy. After a careful selection process in accordance with Institute of Electrical and Electronics Engineers (IEEE) standards, 20 papers were selected for inclusion in the *CyCon 2024 Proceedings*.

Network security is a focus area for the technology track, with **Marc Wyss, Roland Meier, Llorenç Romá, Cyrill Krähenbühl, Adrian Perrig, and Vincent Lenders** examining how to build secure wide-area networks over public internet service providers. **Bruno Dzogovic and Silke Holtmanns** explore securing 5G communication in joint operations by NATO partners. **Pēteris Paikens and Krišjānis Nesenbergs** offer insights into the resilience and vulnerabilities of consumer wireless devices to cyber attacks.

Structured security analysis of complex systems is treated by **Jelizaveta Vakarjuk, Nikita Snetkov, and Peeter Laud**, who identify obstacles relating to post-quantum communication in e-Estonia. **Erwin Orye, Gabor Visky, Alexander Rohl, and Olaf Maennel** analyse the cyber components of sea drones and discuss ways to enhance their cyber resilience.

Several authors address the application of machine learning and artificial intelligence (AI) methods to solve cybersecurity problems. **Yannick Merkli, Roland Meier, Martin Strohmeier, and Vincent Lenders** examined how network flow classifiers could be defeated and improved through adversarial machine learning. **Dan Bogdanov, Paula Etti, Liina Kamm, and Fedor Stomakhin** studied how generalized blueprints can be the basis for an artificial intelligence system risk management methodology. Enhancing military cybersecurity using artificial-intelligence-driven Software Bill of Materials to detect firmware vulnerability and manage assets was examined by **Max Beninger, Philippe Charland, Steven H. H. Ding, and Benjamin C. M. Fung**.

The rapid evolution of emerging technologies such as artificial intelligence (AI) and autonomous weapons is also reshaping international law. The ongoing integration of AI into weapons systems has the potential to significantly impact decision-making and command responsibility in military operations. Legal, policy, and compliance issues regarding the use of AI for security are examined by **Wei-Che Wang**, while **Kristy Chan and Joseph Khaw** investigate the applicability of existing paradigms of international law and interpretation of customary international law to cyberspace. **Jonathan Kwik** analyses divergent interpretations of what constitutes an attack according to international humanitarian law (IHL) and how IHL could be applied to the use of autonomous weapons systems. **Scott Sullivan and Iben Ricket** discuss whether black-box models of AI can consider the nuances of IHL and what aspects policymakers should contemplate when constructing future norms.

The current conflict in Ukraine is and will be a key case study for many, including regarding the unsettled status of collective countermeasures under customary international law. Here, **Jeff Kosseff** examines the international legal framework for Hunt Forwards, while **Lisandra Novo** explores the legality of collective countermeasures in response to malicious cyber operations.

Outer space is an area of growing economic and technological importance. Keeping the peace in outer space has become paramount. **Chris O'Meara** discusses anti-satellite weapons and *jus ad bellum* in the context of self-defence. The potential to harness space debris as a weapon is examined by **Anna Blechovà, Jakub Harašta, and František Kasl**.

Cyber operations can leverage various techniques to deliver effects, which could involve the use of personal data and influencing individuals' perceptions and attitudes, or potentially both. The legal framework for military psychological operations in cyberspace and data protection in times of armed conflict should be considered. **Anastasia Roberts and Adrian Venables** apply a legal framework to military

psychological operations in the digital battlespace. **Tatjana Grote** reflects on the rules governing post-occupation retention and the use of personal data collected by the military.

In the strategy/policy track, two papers address key cross-cutting issues. Now that internet, software, and cybersecurity services have become domains of industry rather than government, **Neil Ashdown** writes about models of public–private cooperation in cyber threat intelligence. **Joseph Jarnecki** addresses the many models and lessons learned from civilian cooperation in rapid assistance to Ukraine which could be modified for future needs.

As in previous years, all articles published in these proceedings have been subject to a double-blind peer review. We are indebted to the members of the CyCon Academic Review Committee, who have taken time out of their already full days to conduct peer reviews and help the Programme Committee in the final selection of papers. We have once again been fortunate to have the continued support of the IEEE and its Estonian branch, without which this volume would not have been possible.

Last, but far from least, the editors would like to thank Jaanika Rannu for her logistical support in the production of these proceedings. A special mention goes to Ingrid Winther, for her work on the strategy track.

Academic Review Committee Members for CyCon 2024:

- Lt. Cmdr. Dr Bernt Åkesson, NATO CCDCOE, Estonia
- Dr Bernhards 'BB' Blumbergs, CERT.LV, Latvia
- Alessandro Boggio Tomasaz, European Union, Belgium
- Sqn. Ldr. Tara Brown, U.S. Naval War College, United States
- Yongkuk Cho, NATO CCDCOE, Estonia
- Dr Sean Costigan, George C. Marshall Centre, Germany
- Sebastian Cymutta, NATO CCDCOE, Estonia
- Dr Samuele De Tomas Colatin, Ca' Foscari University of Venice, Italy
- Dr Talita Dias, Chatham House, United Kingdom
- Lt. (N) Erdi Dönmez, NATO CCDCOE, Estonia
- Dr Andrew Dwyer, Royal Holloway, University of London, United Kingdom
- Dr Helen Eenmaa-Dimitrieva, University of Tartu, Estonia
- Dr Amy Ertan, NATO HQ, Belgium
- Dr Kenneth Geers, Atlantic Council, United States
- Kier Giles, Chatham House, United Kingdom
- Cmdr. Davide Giovannelli, NATO CCDCOE, Estonia

- Dr Andrew Grotto, Stanford University, United States
- Shota Gvineria, Baltic Defence College, Estonia
- Prof. Kimmo Halunen, University of Oulu and National Defence University of Finland, Finland
- Dr Jakub Harašta, Masaryk University, Czech Republic
- Dr Trey Herr, Atlantic Council, United States
- Otakar Horák, NATO CCDCOE, Estonia
- Tat'ána Jančárková, National Cyber and Information Security Agency (NÚKIB), Czech Republic
- Aleksi Kajander, NATO CCDCOE, Estonia
- Dr Agnes Kasper, NATO CCDCOE, Estonia
- Prof. Sokratis Katsikas, Norwegian University of Science and Technology, Norway
- Panagiotis Kirikas, AGT R&D GmbH, Germany
- Dr Csaba Krasznay, National University of Public Service, Hungary
- Erik Kursetgjerde, NATO CCDCOE, Estonia
- Dr Claire Kwan, NATO CCDCOE, Estonia
- Lt. Col. (ret.) Franz Lantenhammer
- Dr Vincent Lenders, armasuisse Science + Technology, Switzerland
- Dr Lauri Lindström, NATO CCDCOE, Estonia
- Dr Erica Lonergan, Columbia University, United States
- Liina Lumiste, University of Tartu, Estonia
- Prof. Kubo Mačák, University of Exeter, United Kingdom
- Matti Mantere, Starship Technologies, Estonia
- Prof. Luigi Martino, University of Florence, Italy
- Dr Paul Maxwell, Army Cyber Institute, United States
- Lt. Cmdr. Michael McCarthy, Canadian Armed Forces, Canada
- Dr Roland Meier, armasuisse Science + Technology, Switzerland
- Dr Stefano Mele, Italian Atlantic Committee, Italy
- Dr Tal Mimran, Hebrew University of Jerusalem, Israel
- Tomáš Minárik, National Cyber and Information Security Agency (NÚKIB), Czech Republic
- Dr Dóra Mólnar, National University of Public Service, Hungary
- Tomomi Moriyama, NATO CCDCOE, Estonia
- Dr Jose Nazario, Mandiant Intelligence, USA
- Lt. Col. Gry-Mona Nordli, Norwegian Armed Forces, Norway
- Dr Piroska Páll-Orosz, Ministry of Defence, Hungary
- Maj. Erwin Oyre, Ministry of Defence, Belgium
- James Pavur, University of Oxford, United Kingdom
- Piret Pernik, NATO CCDCOE, Estonia
- Capt. (N) Jean-Paul Pierini, Italian Navy, Italy

- Col. Peter Pijpers, Ministry of Defence, Netherlands
- Dr Karl Platzer, Austrian Armed Forces, Austria
- Kārlis Podiņš, NATO CCDCOE, Estonia
- Capt. Vasco Prates, NATO CCDCOE, Estonia
- Lt. Col. Graham Price, Australian Defence Forces, Australia
- Dr Matīss Rikters, National Institute of Advanced Industrial Science and Technology (AIST), Japan
- Dr Przemysław Roguski, Jagellonian University, Poland
- Prof. Marco Roscini, University of Westminster, United Kingdom
- Urmas Ruuto, NATO CCDCOE, Estonia
- Kurt Sanger, Integrated Cybersecurity Partners, United States
- Prof. Annita Sciacovelli, University of Bari, Italy
- Lt. Col. Massimiliano Signoretti, Italian Air Force, Italy
- Ben Strickson, Elemendar, United Kingdom
- Dr Johan Sigholm, Swedish Defence University, Sweden
- Dr Zdzislaw Sliwa, Baltic Defence College, Estonia
- Dr Jason Staggs, University of Tulsa, United States
- Dr Tim Stevens, King's College London, United Kingdom
- Siri Strand, King's College London, United Kingdom
- Dr Martin Strohmeier, armasuisse Science + Technology, Switzerland
- Dr Arun Sukumar, Tufts University, United States
- Jens Tölle, Fraunhofer FKIE, Germany
- Grete Toompere, Estonian Defence Forces, Estonia
- Dr Julia Vassileva, Tallinn University, Estonia
- Karine Veersalu, NATO CCDCOE, Estonia
- Dr Adrian Venables, Tallinn University of Technology, Estonia
- Mauro Vignati, International Committee of the Red Cross, Switzerland
- Gábor Visky, Tallinn University of Technology, Estonia
- Prof. Sean Watts, United States Military Academy at West Point, United States
- Dr Laurin Weissinger, Yale Law School, United States
- Cmdr. Michael Widmann, NATO MARCOM, United Kingdom
- Ingrid Winther, Norwegian National Security Authority, Norway
- Lt. Col. Nick Womba, NATO CCDCOE, Estonia
- Dr Jan Wünsche, Swedish Armed Forces, Sweden
- Philippe Zotz, Luxembourg Armed Forces, Luxembourg

CyCon 2024 Programme Committee:

- Dr Sigurður Emil Pálsson, chair
- Dr Claire Kwan, co-chair, editor and track chair (strategy)
- Dr Lauri Lindström, co-chair and editor
- Cmdr. Davide Giovannelli, track chair (law)
- Karlis Podiņš, track chair (technology)
- Lt. Col. Dr Damjan Štrucl, editor

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

On Building Secure Wide-Area Networks over Public Internet Service Providers

Marc Wyss
ETH Zurich
Department of Computer Science
Zurich, Switzerland
marc.wyss@inf.ethz.ch

Roland Meier
armasuisse Science and Technology
Cyber-Defence Campus
Thun, Switzerland
roland.meier@ar.admin.ch

Llorenç Romá
armasuisse Science and Technology
Cyber-Defence Campus
Thun, Switzerland
llorenc.roma@ar.admin.ch

Cyrill Krähenbühl
ETH Zurich
Department of Computer Science
Zurich, Switzerland
cyrill.kraehenbuehl@inf.ethz.ch

Adrian Perrig
ETH Zurich
Department of Computer Science
Zurich, Switzerland
adrian.perrig@inf.ethz.ch

Vincent Lenders
armasuisse Science and Technology
Cyber-Defence Campus
Thun, Switzerland
vincent.lenders@ar.admin.ch

Abstract: Many public and private organizations use wide-area networks (WANs) to connect their geographically distributed sites. Given that these WANs are often critical for the organization's operations, their security with respect to confidentiality, integrity, and availability is crucial.

A high level of security can be reached if the WAN is built with a dedicated network infrastructure, with the organization operating its own layer-2/3 routing, for example, multiprotocol label switching on top of dedicated fibers or leased lines. Unfortunately, this approach is often slow to deploy, requires high operational effort, and is too expensive for many use cases.

A cheaper alternative is to construct the WAN as an overlay network on the infrastructure of public Internet service providers (ISPs), for example, using virtual

private network tunnels between the sites. Unfortunately, the security of such a WAN is suboptimal. For instance, traffic analysis attacks (on encrypted traffic) can reveal sensitive information transmitted over these public networks, compromised routers between the sites can alter packets, and network-layer distributed denial-of-service (DDoS) attacks can disrupt connectivity.

In this paper, we explore a novel inter-ISP network architecture that provides the desired level of control and security for WAN operators, achieving the best of the two above approaches: strong security properties on a cost-efficient public Internet fabric. Our architecture builds on the SCION next-generation Internet architecture and adds extensions for fine-grained path control, connectivity guarantees in the presence of DDoS attacks, and traffic analysis prevention. With this architecture, WAN operators can build on public layer-3 network connectivity services to deploy secure WANs.

Keywords: *wide-area networks (WANs), WAN security, SCION*

1. INTRODUCTION

Wide-area networks (WANs) are often used to connect multiple sites of a large organization. To protect data sent between sites, the WAN must ensure the confidentiality, integrity, and availability (called the CIA triad) of inter-site communication. Ideally, a WAN is built on a dedicated infrastructure consisting of trustworthy network devices and is fully under the control of the organization itself. In practice, WAN deployments frequently make use of leased lines to ensure traffic isolation and to provide quality of service (QoS) guarantees. Unfortunately, building such a WAN requires all sites to be connected via leased lines, which may be prohibitively expensive.

As an alternative, sites can connect to public Internet service providers (ISPs) and use overlay connections over the public Internet which are protected using technologies for virtual private networks (VPNs), such as IPsec tunnels. Such an ISP-based deployment significantly reduces the cost as (1) ISP connectivity typically has a lower cost than leased-line connectivity, and (2) every site only requires a single access point at its upstream ISP. Unfortunately, Internet overlay connections have much weaker security properties compared to leased lines. Thus, they are typically insufficient for the requirements of organizations requiring high security, such as critical infrastructure providers, governmental bodies, and military organizations. For example, data confidentiality is hindered by eavesdropping and traffic hijacking, data integrity may not be preserved due to the presence of adversarial network devices that

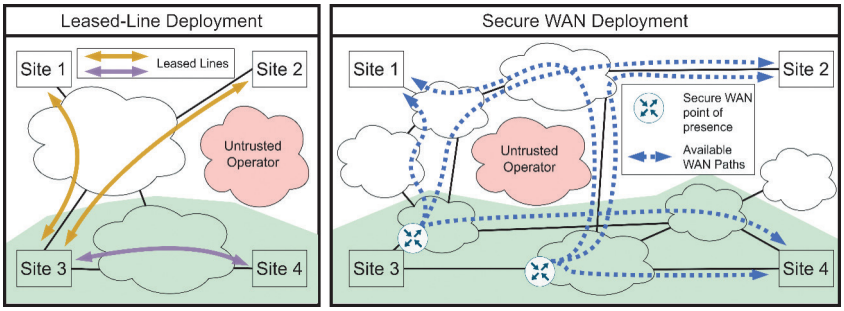
inject or modify packets, and communication availability cannot be guaranteed due to distributed denial-of-service (DDoS) attacks. Furthermore, since the infrastructure is shared, heavy load from other customers can impair the availability of the WAN connections due to network congestion. Centralized solutions, such as software-defined WAN (SD-WAN) [1] approaches and DDoS mitigations based on rerouting and scrubbing [2], can address some of these issues. However, these solutions require a single entity with direct control over the complete network infrastructure, which hampers deployment in widely distributed networks, such as those spanning multiple countries. Furthermore, in a federated setting, finding a single entity trusted by all actors is not always possible.

In this paper, we analyze the challenges of building a secure WAN and propose a WAN architecture that achieves security properties similar to a leased-line deployment—but only requires deploying a single ISP connection per site, resulting in a significant cost reduction. These properties can be achieved by leveraging a recent trend called path-aware networking, which allows endpoints to gain insight into the forwarding paths taken by their traffic (path transparency) and influence the forwarding path choice according to their criteria (path control). Figure 1 compares a leased-line deployment with our proposed secure WAN deployment from the viewpoint of a specific site. In particular, we make use of the next-generation Internet architecture SCION [3], which provides path transparency and control at the level of autonomous systems (ASs), representing individual network operators. Furthermore, we leverage various recent network protocols to achieve the CIA triad, for example, using bandwidth reservations for strict QoS guarantees, DDoS protection, and traffic obfuscation.

Contributions: Our paper makes the following contributions:

- We analyze the security challenges of building a secure WAN over layer-3 connectivity services from ISPs and derive the security properties that an inter-ISP network architecture must support (Section 2).
- We propose an architecture that combines new network protocols to achieve the required security properties (Section 3).
- We present our initial deployment comprising four WAN sites in two countries connected over four ISPs and discuss a potential use case for multinational collaboration (Section 4).

FIGURE 1: DEPLOYMENT SCENARIOS FOR SECURELY CONNECTING MULTIPLE SITES FROM THE PERSPECTIVE OF SITE 3 (CONNECTING TO ALL OTHER SITES). IN THE FIRST SCENARIO, THE SITES ARE CONNECTED VIA PAIRWISE LEASED-LINE CONNECTIONS, AND IN THE SECOND SCENARIO VIA OUR PROPOSED SECURE WAN SOLUTION. TRAFFIC BETWEEN SITE 3 AND SITE 4 SHOULD NOT LEAVE A CERTAIN JURISDICTION (GREEN AREA)—FOR EXAMPLE, TO PRESERVE POLICY COMPLIANCE—AND ALL TRAFFIC SHOULD AVOID A CERTAIN UNTRUSTED NETWORK OPERATOR (RED)



2. SECURITY CHALLENGES

Our aim is to craft a secure WAN solution tailored to the public Internet. In this context, "secure" encompasses the CIA triad—confidentiality, integrity, and availability—which is crucial for WANs: Confidentiality in the WAN context means safeguarding all traffic, both metadata and payload, and minimizing data proliferation. Integrity involves ensuring that data remains unaltered during transmission and enabling the detection of unauthorized tampering. Availability ensures seamless communication, allowing successful transmission of all traffic up to a specific bandwidth threshold between all WAN sites.

Table I highlights significant threats against those security properties and presents possible mitigation measures. We do not claim that our identification of threats is complete, but the threats discussed here represent our best estimate of the most significant challenges.

TABLE I: THREATS AGAINST THE CONFIDENTIALITY, INTEGRITY, AND AVAILABILITY OF WAN COMMUNICATION AND POSSIBLE MITIGATION MEASURES. SOME MITIGATION MEASURES CAN HAVE THE SIDE EFFECTS OF CONTRIBUTING TO MITIGATING AND/OR AMPLIFYING CERTAIN THREATS

	Mitigations → / Threats ↓	Traffic encryption	Traffic shaping and padding	Traffic filtering	Traffic authentication	Path authentication	Path control	Traffic prioritization
Confidentiality	Eavesdropping (payloads)	V					V/X	
	Eavesdropping (metadata)	V	V				V/X	
	Traffic hijacking					V	V	
Integrity	Traffic injection				V			
	Traffic modification				V			
Availability	Traffic dropping			X			V	
	Traffic hijacking					V	V	
	Congestion		X				V/X	V
	Volumetric DDoS			V	V		V	V
	Topology changes						V	

V Mitigates the threat
V Can contribute to mitigating the threat
X Can contribute to amplifying the threat in general
X Can contribute to amplifying the threat in general but does not apply to the architecture proposed in Section 3

A. Threats

Let us now discuss the threats listed in Table I in more detail.

1) Confidentiality

Confidentiality on the Internet covers both the secrecy of communication and the requirement of network operators to hide sensitive parts of their network topologies. *Eavesdropping*, for example, involves unauthorized interception of data in the network, potentially compromising confidentiality by allowing malicious entities to access sensitive information without the knowledge of the sender or receiver. If traffic is encrypted, an adversary cannot directly infer sensitive information from the packets' payload, but even eavesdropping on encrypted traffic can be problematic, as it can reveal metadata that enables traffic analysis.

Traffic analysis concerns the analysis of patterns, behaviors, or characteristics within data transmission, posing a risk to confidentiality by allowing the inference of sensitive information, even if traffic is encrypted and source or destination addresses are unknown. Traffic analysis can, for example, infer visited websites, streamed videos, or voice over IP (VoIP) calls from traffic volume, packet sizes, or timing information [4]. Furthermore, as an emerging technology, quantum computing presents a looming threat to current cryptographic algorithms. It has the potential to compromise the confidentiality of previously recorded encrypted traffic by breaking encryption algorithms that are currently considered secure.

Traffic hijacking allows an off-path adversary to redirect traffic across its network nodes, to essentially become an on-path adversary and gain access to confidential communication flows. An adversary can then eavesdrop and perform traffic analysis. This enables powerful eavesdropping attacks as it significantly increases the attack surface.

2) Integrity

Traffic injection, commonly referred to as packet spoofing, involves creating and transmitting network packets with falsified source addresses or other misleading information. It jeopardizes integrity by potentially circumventing defense systems, allowing for the exploitation of vulnerabilities, and enabling many different types of denial-of-service attacks. Additionally, control messages such as from the Internet control message protocol can be spoofed, leading to misinformation and potentially disrupting network operations.

Traffic modification refers to the unauthorized alteration of data packets during transmission, directly compromising integrity. Packets may be modified due to various factors such as transmission errors, malicious manipulation, or faults within packet processing systems. A common attack that leverages traffic modification is a man-in-the-middle attack, where an adversary places itself between the communicating endpoints and modifies selected packets. Quantum computing may impact traffic integrity by potentially breaking asymmetric cryptography used in digital signatures.

3) Availability

Achieving availability is often considered more challenging than confidentiality or integrity [5].

Traffic drop, such as where an on-path attacker deliberately drops either selected or simply all packets, is one of the most difficult threats to mitigate.

Path hijacking involves malicious entities diverting network traffic away from its intended path. It impacts availability by rerouting traffic through unauthorized paths, leading to potential delays or packet loss. Off-path attackers often use path hijacking attacks to essentially become on-path attackers. Path hijacking attacks are particularly powerful in combination with traffic-dropping attacks.

Congestion, whether naturally occurring due to high network demand or because of volumetric DDoS attacks, can affect both network and server availability. Network congestion leads to delays, packet loss, and decreased throughput, hindering communication, while DDoS attacks against servers can render their hosted services inaccessible to legitimate users. Volumetric DDoS attacks are a major reason why WAN operators prefer using leased lines to communicating over the public Internet, a trend amplified by the fact that attacks have grown in strength over recent years [6], [7]. Well-connected entities with links at hundreds of gigabits of capacity, as well as distributed botnets, sometimes leveraging powerful cloud virtual machines, can execute such attacks at a large scale [8]. Additionally, the proliferation of 10 Gbps home connections has made volumetric DDoS attacks easier and even more potent.

Topology changes regarding the network's physical or logical structure can potentially disrupt established communication paths. Such changes not only comprise topology modifications such as rewiring but also link or router failures. Link failures, whether due to hardware issues or physical damage, compromise all communications that use the link as part of their forwarding path. Router failures, whether caused by power outages, hardware malfunctions, natural disasters, or malfunctions or bugs in networking software, similarly affect availability. In such cases, data flows might be interrupted or rerouted; however, routing protocols may take time to converge and find the best available paths, as routers need to adjust and reach a consistent view of the network. During this convergence period, there can be temporary inconsistencies in routing information resulting in poor QoS, including high latency or jitter, and even lost connectivity, rendering services communicating over the WAN unusable.

B. Mitigation Measures

In this section, we discuss some possible mitigation measures for threats to a secure WAN architecture.

1) Traffic Encryption

Implementing encryption—for example, using transport layer security or IPsec—ensures that data transmitted across the network remains illegible to unauthorized entities. Also, employing multiple independent layers of encryption can effectively diminish the risk associated with misconfigurations and vulnerabilities. In light of the

impending threat posed by quantum computing, the adoption of post-quantum secure encryption becomes paramount.

2) Traffic Shaping and Padding

Traffic shaping involves regulating the flow of data toward a consistent and controlled transmission rate. Smoothing out data bursts or adding chaff packets can make traffic patterns less detectable. Thus, it becomes harder for an eavesdropper to discern specific patterns or extract meaningful information from a shaped traffic flow. In particular, shaping obscures the timing and volume of data transmission, thereby making it more challenging for adversaries to determine the nature and content of the communication. Padding involves adding extra data to packets to obscure the actual size or structure of the transmitted packets, rendering eavesdropping less effective. While shaping may delay the delivery of data, padding introduces additional data that occupies network bandwidth without conveying useful information and thus has the side effect of wasting resources.

3) Traffic Filtering

Network devices apply traffic filtering on different layers. One example is a firewall, which may filter packets at the transport and network layer or perform deep packet inspection to check application-layer data. By creating a boundary between a network and the Internet, a firewall can effectively protect against external network scanning while still allowing the network operator to debug the network from inside. Furthermore, firewalls can protect services hosted within the network from outside adversaries by rate-limiting incoming requests. Traffic filtering is also used for defensive mechanisms at end hosts that safeguard against DDoS attacks, ensuring that benign traffic can always successfully reach the destination application without being dropped due to congestion at the receiving end. Depending on the filtering technique, traffic filtering can have a negative impact on network availability if traffic is dropped by mistake (in false positive classifications).

4) Traffic and Path Authentication

The authentication of traffic is important both in the control plane and in the data plane.

Authentication in the control plane ensures that adversaries cannot tamper with control plane messages, such as route announcements sent by ASs and address prefix authorization messages issued by a public key infrastructure (PKI), thus preventing path hijacking attacks. Additionally, if control messages sent to endpoints are authenticated, an adversary cannot produce fake control messages, which may be used to revoke existing paths or inject non-existing devices in a traceroute reply.

Finally, by authenticating control messages originating at endpoints, a firewall can reply to control messages from authorized endpoints only, allowing networks to hide topology information.

In the data plane, authentication is used to ensure the integrity of end-to-end traffic, where authentic keys are typically fetched from a trusted PKI. Source authentication also prevents many different types of DDoS attacks that rely on spoofed source addresses, for example, reflection and amplification attacks, because firewalls can drop unauthenticated traffic.

5) Path Control

Path control, requiring a certain level of path transparency, ensures that traffic follows predefined routes, preventing it from leaving designated areas or regions. From an organization's WAN perspective, fine-grained path transparency and control are desirable because this enables traffic steering through trusted routers. Furthermore, path control mitigates path hijacking, as an attacker cannot influence the forwarding path through routing attacks. It also enables routing traffic around failing or congested links and routers, and around infrastructure believed to eavesdrop and perform traffic analysis. In addition to selecting preferred paths, the network should also provide path validation, that is, a mechanism to ensure that traffic is indeed sent over the selected path. Path validation can further be strengthened by remote attestation of on-path routers to ensure that they exist, are correctly configured, and run the intended software. Finally, path control can mitigate the threat of quantum computing by circumventing an adversary such that it cannot record and later decrypt the data.

In addition to selecting a desirable path to send traffic, endpoints can also leverage path control to enable resource pooling and simultaneously send data across multiple paths. This allows endpoints to achieve higher throughput but comes with its own challenges. The sender must ensure that no adversary is located on any of the selected paths to protect against eavesdropping and that other non-multipath flows do not experience excessive congestion, for example, using fair multipath congestion control protocols [9].

6) Traffic Prioritization

Traffic prioritization improves the QoS of priority traffic by instructing on-path routers to forward it with a higher priority compared to other traffic. This allows network operators to support low-latency and low-jitter communication while simultaneously maximizing the overall link usage for latency-insensitive best-effort traffic, such as file transfers.

A specific application of traffic prioritization is to explicitly reserve the required bandwidth for the priority traffic. Allocating bandwidth along the forwarding path between two WAN sites using an inter-domain bandwidth reservation protocol [10] can ensure that critical data flows smoothly through the network, even during periods of high demand or congestion. As traffic sent over a bandwidth reservation is unaffected by congestion, it shows properties similar to leased lines.

3. PROPOSED ARCHITECTURE FOR SECURE WANS

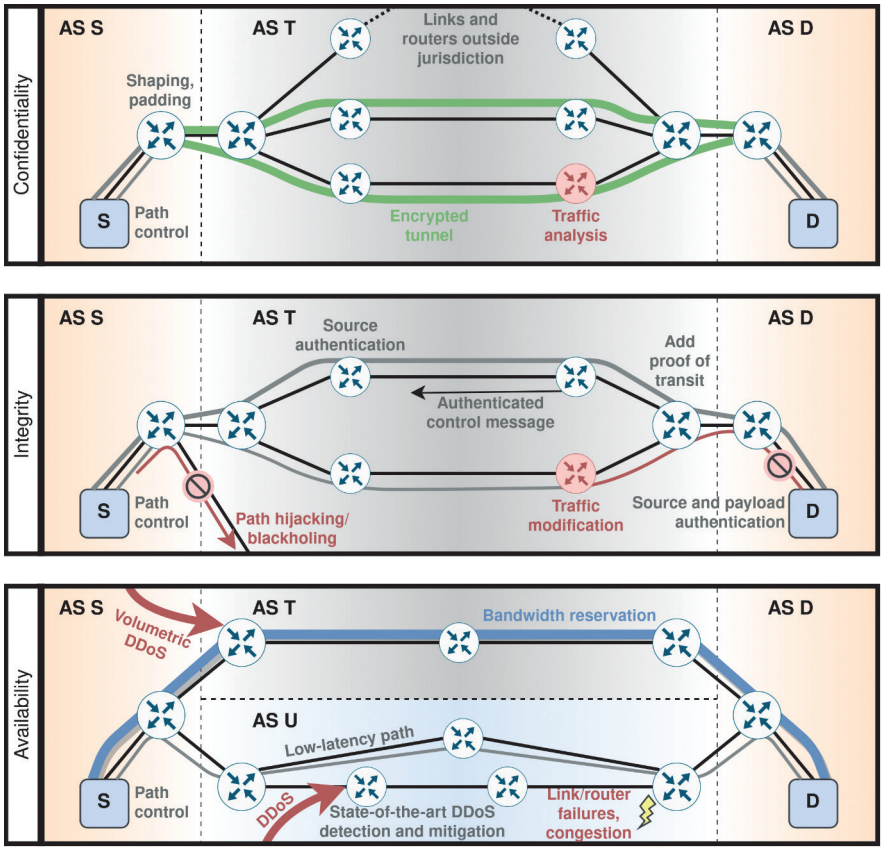
Considering the threats and the methods of mitigating them outlined earlier, we present an architecture that integrates a concise selection of mitigation measures optimized for WAN operators. Our goal is to streamline deployment complexity at ISPs by minimizing the number of mitigation measures while ensuring comprehensive coverage against the identified threats. While the selected mitigation measures address a wide range of threats, our architecture remains flexible, allowing for the inclusion of supplementary measures as needed.

A. Overview

The core component of our architecture is the next-generation Internet architecture SCION [3], which provides AS-level path control and authenticated control messages. SCION serves as the foundation for several systems: a bandwidth reservation protocol (called Helia [10]), a protocol to increase the path selection granularity to individual routers (called FABRID [11]), and a mechanism to guarantee connectivity to end hosts and services despite volumetric DDoS attacks (called Lightning Filter [3]). To ensure the secrecy of the transmitted data, the communication between sites is encrypted using VPN tunnels. In addition, we leverage a state-of-the-art DDoS defense solution called ACC-Turbo [12] to protect against pulse-wave DDoS attacks. Finally, we further improve privacy through traffic shaping and policing offered by a system called DITTO [13].

Figure 2 illustrates our solution, and in the following subsections, we provide more details about each of these components.

FIGURE 2: ILLUSTRATION OF VARIOUS THREATS AGAINST CONFIDENTIALITY, INTEGRITY, AND AVAILABILITY OF COMMUNICATIONS BETWEEN A SOURCE S AND A DESTINATION D IN A WAN DEPLOYMENT OVER PUBLIC INTERNET INFRASTRUCTURE, AND THE SOLUTIONS OUR ARCHITECTURE INCORPORATES TO MITIGATE THEM



B. SCION as the Foundation

SCION is a next-generation Internet architecture designed to improve security, scalability, and control over network traffic. SCION achieves path transparency by making the available (AS-level) forwarding paths visible to end hosts and allows path control by enabling end hosts to select among a set of offered paths. This improves reliability by allowing end hosts to quickly transition to alternative paths and can mitigate congestion through in-network multipath. SCION mitigates path hijacking attacks by design, since routing information is cryptographically secured in both the control and the data plane. Furthermore, control messages such as traceroute replies are authenticated to prevent off-path adversaries from modifying or injecting control messages. Based on the SCION architecture, researchers have developed several

mechanisms for more fine-grained path transparency and control as well as mechanisms to achieve minimum communication guarantees despite volumetric DDoS attacks. We leverage some of these extensions for our proposed WAN architecture. In the following sections, we provide more details about them.

C. Traffic Encryption with IPsec

IPsec [14] is a widely used protocol suite for encrypting and authenticating network traffic. IPsec can be used, among other things, to create encrypted and authenticated layer-3 tunnels between two endpoints. For these tunnels, the original IP packets are encrypted and encapsulated in a new IP packet. Therefore, only the IP addresses of the tunnel's endpoints are visible to potential eavesdroppers. We rely on IPsec tunnels to encrypt communications between WAN sites.

D. Intra-Domain Path Selection, Packet Source Authentication, and Path Validation with FABRID

FABRID extends SCION's capabilities of path transparency and control to the intra-domain router level. It is the first system that enables applications, and hence WAN operators, to forward traffic flexibly, potentially on multiple paths, selected to comply with user-defined preferences [11]. In FABRID, network operators communicate information about their internal router topologies in the form of router policies. These policies are then made accessible to applications along with the set of selectable forwarding paths. This allows each application to choose suitable router policies and encode the chosen policies within its data packets. Consequently, routers along these paths understand how to route traffic in accordance with the designated policies. These policies can encompass diverse router attributes within an AS, such as hardware specifications, geographic location, or manufacturer details. For instance, sensitive data may have to be sent exclusively through routers within specific jurisdictions. Likewise, services reliant on precise time synchronization might need the exclusive use of precision time protocol-capable routers. Some entities, such as governments or critical infrastructure operators, may mandate that traffic should avoid routers with known vulnerabilities or routers produced by untrusted manufacturers, resulting in paths comprised exclusively of recognized, trustworthy equipment. Apart from fine-tuning path transparency and control, FABRID enables on-path routers to cryptographically authenticate the source of each packet and extend the packets with proofs of transit, providing path validation for the source and destination hosts.

E. Bandwidth Reservation with Helia

Helia is a secure inter-domain bandwidth reservation protocol, allowing the dynamic allocation of bandwidth over SCION. It builds on the concept of flyover reservations, a fundamentally new approach for addressing the availability demands of critical low-volume applications. In contrast to path-based reservation systems, flyovers

are fine-grained "hop-based" bandwidth reservations on the level of individual ASs [10]. As Helia can offer forwarding guarantees despite large-scale volumetric DDoS attacks against network infrastructure, such that no off-path adversary can prevent the successful delivery of traffic sent over the reservation, Helia can be regarded as a cost-efficient alternative to the inherent guarantees of leased lines.

F. DDoS Protection with Lightning Filter and ACC-Turbo

DDoS attacks can target end hosts (e.g., web servers) or the network infrastructure (links or routers). For each of these targets, we propose a suitable defense system.

Lightning Filter [3] is a high-speed filtering system that protects (groups of) hosts or services from volumetric DDoS attacks. Just as Helia guarantees the successful forwarding of traffic through the network, Lightning Filter ensures access to the protected destination. Lightning Filter prevents DDoS attacks that rely on spoofed source IP addresses because it authenticates the source and payload of all incoming packets. Since currently deployed firewalls are often susceptible to such attacks, they can profit from Lightning Filter as a first layer of defense. Lightning Filter can thus reduce the packet drop rate under DDoS attacks by reducing the amount of traffic volume that has to be processed by specialized firewalls, and it can improve the accuracy of these firewalls by removing spoofed packets.

Volumetric DDoS attacks against routers and network links can be mitigated through bandwidth reservations, which provide fundamental availability guarantees. However, not all traffic is equally important, and therefore, a portion of traffic might be sent as best-effort traffic without bandwidth reservations. We therefore rely on ACC-Turbo [12] as an additional DDoS defense system. ACC-Turbo is highly effective in detecting and mitigating even pulse-wave DDoS attacks at line rate, making it the fastest in-network DDoS mitigation technique to date. Such in-network DDoS defense systems are significantly more effective in the absence of spoofed source addresses and thus benefit from FABRID's source authentication.

G. Traffic Analysis Prevention with DITTO

Even though traffic is encrypted with IPsec tunnels between the WAN sites in our architecture, communication is still vulnerable to traffic analysis attacks. While FABRID's fine-grained path control can mitigate such attacks to some degree, as it makes it possible to steer traffic around untrusted devices, this measure is insufficient if trust is not warranted. Furthermore, traffic analysis might occur not only at malicious or compromised devices but also at network links. Compared to on-path attackers actively delaying or dropping packets, detecting attackers performing traffic analysis is nearly impossible. We therefore rely on DITTO [13], a traffic obfuscation system to protect against traffic analysis. DITTO has been specifically designed to protect

links to WAN sites, scaling up to 100 Gbps links. In addition, it does not require modifications at the end hosts. DITTO adds padding and chaff packets at line rate, ensuring that outgoing traffic always follows a fixed pattern.

H. Summary

A combination of the abovementioned components results in a cost-efficient architecture to securely operate WANs over the public Internet. As summarized in Table II, this architecture implements comprehensive mitigations against all the threats discussed in Section 2. However, it is essential to note the potential drawback that traffic shaping and padding may increase the risk of congestion (Table I).

TABLE II: COMPONENTS OF OUR PROPOSED WAN ARCHITECTURE AND THE MITIGATIONS THEY IMPLEMENT

	Traffic encryption	Traffic shaping and padding	Traffic filtering	Traffic authentication	Path authentication	Path control	Traffic prioritization
IPsec	V			V			
DITTO		V					
Lightning Filter			V	V			V
FABRID				V		V	
Helia				V			V
SCION					V	V	
ACC-Turbo							V

4. THE ROAD TO DEPLOYMENT

Given that many of the components mentioned in Section 3 are still research prototypes, there is a gap between our proposed architecture and current ISP service offerings. To evaluate our architecture, we deployed a testbed across four WAN sites. In this section, we first describe our testbed and then focus on the technological readiness of each individual component. Afterwards, we discuss the remaining challenges and outline a possible use case.

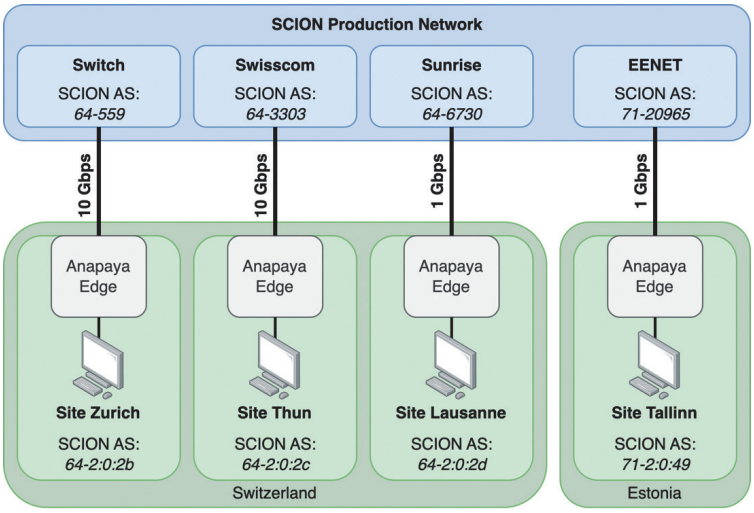
A. Testbed Deployment

We created a testbed that uses the SCION infrastructure provided by four different ISPs to establish a secure WAN connecting multiple locations. This deployment is primarily for experimental use, rather than incorporation into a production environment. Initially, we connected three separate locations across Switzerland to the SCION network over three different ISPs: Swisscom, Sunrise, and Switch. We expanded this deployment with an additional WAN node in Tallinn (Estonia) with EENet as the ISP. Each of these locations forms its own SCION AS.

Our setup relies on Anapaya Edge [15], a commercially available device running a SCION-IP gateway (SIG) and a SCION router. Anapaya is a spin-off of ETH Zurich, where large parts of SCION were developed. The SIG acts as a tunnel endpoint for IP packets, encapsulating them into SCION packets and delivering them to the specified SCION addresses. This way, end hosts and applications in our WAN deployment do not need to support SCION, as the SIG implements the whole protocol translation and path selection logic. Link access to the SCION network is provided through distinct ISPs that have established direct peering among each other.

In this SCION setup, all participants were assigned one or multiple AS identifiers, and the corresponding list is publicly available [16]. Figure 3 shows an overview of our current deployment.

FIGURE 3: OVERVIEW OF OUR SCION-BASED WAN DEPLOYMENT CONNECTING FOUR SITES IN TWO COUNTRIES



Our deployment allows us to direct all traffic between the sites natively over the SCION network. A major advantage of SCION is the availability of many paths, enabling the selection of various routes for routing traffic through a series of distinct hops. Figure 4 shows a subset of eight existing paths from the Anapaya Edge in Zurich to the Anapaya Edge in Thun. While SCION offers different AS-level paths, we can observe that there are also paths that traverse the same ASs and differ only with respect to their ingress and egress interfaces.

FIGURE 4: SUBSET OF AVAILABLE FORWARDING PATHS FROM ZURICH TO THUN AS SHOWN AT THE SIG IN ZURICH. PATHS ARE REPRESENTED AS AUTOMATIC SYSTEM SEQUENCES (HOPS) AND INTERFACE PAIRS. AN INTERFACE PAIR IS REPRESENTED AS "EG>IN," WHERE "EG" IS THE EGRESS INTERFACE ID AND "IN" IS THE INGRESS INTERFACE ID

```
Available paths to 64-2:0:2c
4 Hops:
[0] Hops: [64-2:0:2b 1>24 64-559 17>1 64-3303 21>1 64-2:0:2c]
[1] Hops: [64-2:0:2b 1>24 64-559 17>1 64-3303 25>2 64-2:0:2c]
[2] Hops: [64-2:0:2b 1>24 64-559 19>9 64-3303 21>1 64-2:0:2c]
[3] Hops: [64-2:0:2b 1>24 64-559 19>9 64-3303 25>2 64-2:0:2c]
5 Hops:
[4] Hops: [64-2:0:2b 1>24 64-559 4>15 64-2:0:13 18>4 64-3303 21>1 64-2:0:2c]
[5] Hops: [64-2:0:2b 1>24 64-559 4>15 64-2:0:13 18>4 64-3303 25>2 64-2:0:2c]
[6] Hops: [64-2:0:2b 1>24 64-559 16>10 64-6730 11>11 64-3303 21>1 64-2:0:2c]
[7] Hops: [64-2:0:2b 1>24 64-559 16>10 64-6730 11>11 64-3303 25>2 64-2:0:2c]
[...]
```

We leveraged this level of path transparency to find all inter-domain paths offered by SCION at site Zurich. Table III shows the result of this evaluation: the total number of available paths, including paths that only differ in terms of their interfaces, from Zurich to the other sites in our deployment.

TABLE III: NUMBER OF AVAILABLE PATHS FROM ZURICH TO THUN, LAUSANNE, AND TALLINN. THE COLUMNS REPRESENT THE PATH LENGTH IN TERMS OF AS-LEVEL HOPS

	Hops			
Destination	**4**	**5**	**6**	**7**
Thun	4	32	66	8
Lausanne	2	14	35	-
Tallinn	2	-	-	-

We first verified the general connectivity between the sites using the SCION ping utility [17] without specifying a forwarding path, meaning that the SIG automatically

selected the path to each destination. The resulting round-trip time (RTT) and jitter are shown in Table IV. As expected, the RTT between sites in the same country was significantly lower than the RTT to the remote site in Tallinn.

TABLE IV: RTT AVERAGE AND STANDARD DEVIATION IN MILLISECONDS FOR 20 PROBE PACKETS FROM ZURICH AND TALLINN TO ALL OTHER WAN SITES. NOTE THAT THESE RESULTS DEPEND ON THE PATH SELECTED BY THE SIG, WHICH MAY CHANGE OVER TIME

	Destination			
Source	**Zurich**	**Thun**	**Lausanne**	**Tallinn**
Zurich	-	12.08 (0.05)	4.89 (0.09)	59.42 (0.15)
Tallinn	60.09 (0.41)	72.77 (0.22)	65.12 (0.33)	-

To evaluate the impact of different inter-domain paths on the RTT, we measured the RTT between Zurich and Thun using various paths of different lengths. This time, we used the SCION ping utility to explicitly specify the forwarding path. We manually selected a total of 12 paths, three each of lengths four, five, six, and seven. Table V shows the obtained RTT and jitter values. The measurements show that paths of the same length can vary significantly in terms of their RTT. At the same time, longer paths do not necessarily imply a higher RTT, as can be observed for the values up to six AS-level hops, where the RTT values are similar. We observe consistently low jitter irrespective of the forwarding path.

TABLE V: RTT AVERAGE AND STANDARD DEVIATION IN MILLISECONDS FROM ZURICH TO THUN FOR PATHS WITH DIFFERENT NUMBERS OF AS-LEVEL HOPS, AVERAGING OVER 20 PROBES. EACH COLUMN REFERS TO A DIFFERENT PATH

	4 hops			5 hops			6 hops			7 hops		
RTT (ms)	5.4	9.3	11.7	5.7	9.5	12.0	6.2	9.6	15.3	11.7	18.6	22.1
RTT standard deviation (ms)	0.04	0.08	0.06	0.44	0.03	0.07	0.06	0.52	0.09	0.07	1.10	0.19

B. Readiness of the Proposed Components

Technology readiness level (TRL) is a widely used metric to describe the maturity of a technology [18]. It uses a scale from TRL 1, which applies to technology whose basic principles have been observed and reported, to TRL 9, which applies to technology that is used in an actual system in production.

For each component of our proposed architecture, Table VI indicates the current TRL and whether there exist ISPs that readily offer it as a service. Most technologies are in a rather early stage (TRL 3)—this is because they were only recently proposed by academic research. For some components, ISP support is not needed because they can be implemented end-to-end by the WAN operator without any ISP integration. SCION is fully supported by the four commercial ISPs in our testbed and offered as an experimental service in their operational environment (TRL 7).

TABLE VI: TRLS OF THE COMPONENTS COMPRISING OUR PROPOSED WAN ARCHITECTURE

Technology	Offered by ISPs	TRL
IPsec	Not needed	9 (actual system proven in operational environment)
SCION connectivity	Yes	7 (system prototype demonstration in operational environment)
FABRID	Not yet	3 (experimental proof of concept)
Helia	Not yet	3
Lightning Filter	Not yet	3
ACC-Turbo	Not yet	3
DITTO	Not needed	3

C. Remaining Challenges

There are several challenges in implementing and deploying our proposed WAN architecture. The biggest of these is the lack of support from commercial ISPs for the needed security components (see Table VI). The exception is SCION connectivity, which is already provided by commercial ISPs. However, SCION is not yet available in all countries, meaning that it may not be possible to connect some WAN sites to the production network over an arbitrary ISP. Still, the current SCION deployment readily covers networks in Europe, Asia, and North America and is rapidly expanding [16]. There are also challenges regarding differing hardware requirements; the different security solutions have only been evaluated independently of each other so far. ACC-Turbo and DITTO have been implemented in P4-compatible devices. Today, these devices do not natively support cryptographic algorithms such as Advanced Encryption Standard (AES). However, AES is required by SCION and its extensions. Recently, a P4 router implementation has been presented that uses an accelerator for the cryptographic validations and can thus forward SCION traffic on the Intel Tofino 2 [19] at a rate of more than 3 terabits per second [20]. Given these advancements, a P4 implementation of Helia, FABRID, and Lightning Filter, each requiring further

AES key expansions and block cipher computations compared to SCION, might soon become feasible. Nevertheless, the components of our proposed WAN solution can also be deployed on other platforms; for example, some components have been implemented and evaluated using Intel's Data Plane Development Kit (DPDK). In financial terms, precisely estimating or predicting the costs of our proposed architecture is challenging. Nevertheless, we expect the costs to be significantly lower compared to solutions such as leased lines, as our architecture operates on an existing, cost-efficient public Internet fabric and existing SCION routers used in production rely on DPDK, therefore new security mechanisms can be installed through software updates. Lastly, while every component has already been independently analyzed for its security properties, the security of the entire system has not been explored—an interesting and important avenue for future work.

D. Use Case: Multinational Collaboration

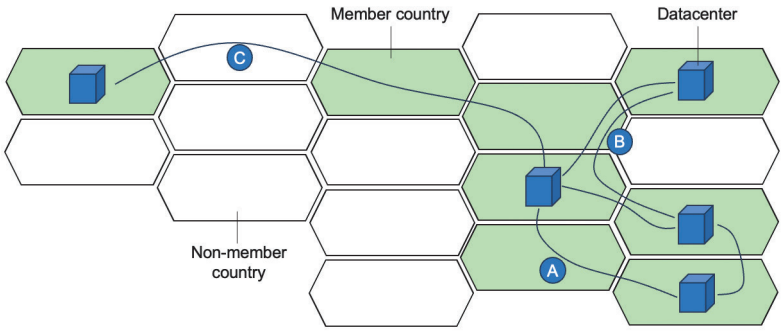
Our architecture is suitable for many types of WANs, from small companies to large multinational organizations with sites across the globe. Besides our own deployment, another interesting use case could be secure communications among different countries, for example, transmitting sensitive data between their data centers, where this traffic should ideally only traverse other member countries.

As such, this scenario has similarities with the collaboration among NATO countries, which might want to securely interconnect their data centers or cyber ranges for training purposes [21].

As illustrated in Figure 5, our proposed architecture constitutes an ideal solution for the following reasons:

- The individual sites can be operated independently by the respective countries.
- WAN connections use existing infrastructure; therefore, no additional cables need to be installed.
- Path control makes it possible to ensure that traffic only crosses member countries, even if it is not the shortest path.
- If there is no path that traverses only member countries, the number of non-member countries can be minimized. Encryption and traffic shaping minimize the attack surface in this case.

FIGURE 5: WAN CONNECTING DATA CENTERS LOCATED IN DIFFERENT COUNTRIES. OUR ARCHITECTURE ALLOWS ROUTING TRAFFIC ONLY ALONG PATHS THAT DO NOT LEAVE MEMBER COUNTRIES IF THIS IS POSSIBLE (A, B). IF IT IS NOT POSSIBLE (C), THE NUMBER OF TRAVERSED NON-MEMBER COUNTRIES CAN BE MINIMIZED AND OTHER MITIGATION MEASURES LIMIT THE RISK OF A SUCCESSFUL ATTACK



5. CONCLUSION

In this paper, we analyzed the challenges in designing secure WANs for large organizations to interconnect their geographically distributed sites. We found that such a WAN can be built on the shared infrastructure of public ISPs thanks to the increasing adoption of the SCION next-generation Internet architecture and recently introduced technologies. Our proposed solution leverages SCION and recent results from the research community to achieve strong security guarantees while significantly reducing costs compared to the currently used WAN approaches that are based on leased lines.

To verify the feasibility of this approach, we implemented and evaluated basic SCION connectivity at multiple WAN sites in two countries. In future work, we plan to extend this testbed both geographically by adding more sites and technologically by implementing and deploying the missing components.

Once comprehensively implemented and deployed, the proposed architecture would allow organizations of any size to build secure WANs over the public Internet.

REFERENCES

[1] Z. Yang, Y. Cui, B. Li, Y. Liu, and Y. Xu, "Software-defined wide area network (SD-WAN): Architecture, advances and opportunities," in *Proceedings of the International Conference on Computer Communications and Networks (ICCCN)*, Jul. 2019, doi: 10.1109/icccn.2019.8847124.

[2] P. Zilberman, R. Puzis, and Y. Elovici, “On network footprint of traffic inspection and filtering at global scrubbing centers,” *IEEE Trans. Dependable Secure Comput*., vol. 14, no. 5, 2015.
[3] L. Chuat et al., *The Complete Guide to SCION*. Springer International Publishing, 2022, doi: 10.1007/978-3-031-05288-0.
[4] E. Papadogiannaki and S. Ioannidis, “A survey on encrypted network traffic analysis applications, techniques, and countermeasures,” *ACM Comput. Surv*., vol. 54, no. 6, Jul. 2021, doi: 10.1145/3457904.
[5] G. Schmid, “Thirty years of DNS insecurity: Current issues and perspectives,” *IEEE Commun. Surv. Tutor*., vol. 23, no. 4, 2021, doi: 10.1109/comst.2021.3105741.
[6] “DDoS threat intelligence report: Issue 11.” NETSCOUT. Dec. 2023. [Online]. Available: https://perma.cc/V44P-543K
[7] O. Yoachimik. “Cloudflare DDoS threat report for 2022 Q4.” Cloudflare Blog. Oct. 2023. [Online]. Available: https://perma.cc/Q6VB-BG2J
[8] “DDoS threat report for 2023 Q3.” Cloudflare Blog. Accessed: Jan. 4, 2024. [Online]. Available: https://blog.cloudflare.com/ddos-threat-report-2023-q3
[9] C. Xu, J. Zhao, and G.-M. Muntean, “Congestion control design for multipath transport protocols: A survey,” *IEEE Commun. Surv. Tutor*., vol. 18, no. 4, 2016, doi: 10.1109/comst.2016.2558818.
[10] M. Wyss, G. Giuliari, J. Mohler, and A. Perrig, “Protecting critical inter-domain communication through flyover reservations,” in *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, Nov. 2022, doi: 10.1145/3548606.3560582.
[11] C. Krähenbühl, M. Wyss, D. Basin, V. Lenders, A. Perrig, and M. Strohmeier, “FABRID: Flexible attestation-based routing for inter-domain networks,” in *Proceedings of the USENIX Security Symposium*, 2023.
[12] A. G. Alcoz, M. Strohmeier, V. Lenders, and L. Vanbever, “Aggregate-based congestion control for pulse-wave DDoS defense,” in *Proceedings of the ACM SIGCOMM Conference*, 2022, doi: 10.1145/3544216.3544263.
[13] R. Meier, V. Lenders, and L. Vanbever, “Ditto: WAN traffic obfuscation at line rate,” in *Proceedings of the Symposium on Network and Distributed Systems Security (NDSS)*, 2022, doi: 10.14722/ndss.2022.24056.
[14] “IP security protocol (IPsec).” IETF Datatracker. Accessed: Dec. 29, 2023. [Online]. Available: https://datatracker.ietf.org/wg/ipsec/
[15] “Anapaya Edge overview.” Anapaya. Accessed: Apr. 4, 2024. [Online]. Available: https://docs.anapaya.net/en/latest/edge/overview/
[16] “ISD and AS assignments.” Anapaya. Accessed: Jan. 4, 2024. [Online]. Available: https://docs.anapaya.net/en/latest/resources/isd-as-assignments/
[17] “SCION ping.” Anapaya. Accessed: Jan. 5, 2024. [Online]. Available: https://scion.docs.anapaya.net/en/latest/command/scion/scion_ping.html
[18] “Technology readiness levels.” NASA. Accessed: Jan. 5, 2024. [Online]. Available: https://www.nasa.gov/directorates/somd/space-communications-navigation-program/technology-readiness-levels/
[19] “Intel Tofino 2.” Intel. Accessed: Jan. 4, 2024. [Online]. Available: https://www.intel.com/content/www/us/en/products/details/network-io/intelligent-fabric-processors/tofino-2.html
[20] L.-C. Schulz, R. Wehner, and D. Hausheer, “Cryptographic path validation for SCION in P4,” in *Proceedings of the 6th on European P4 Workshop*, 2023.
[21] “Cyber Ranges Federation Project reaches new milestone.” European Defence Agency. Accessed: Jan. 4, 2024. [Online]. Available: https://eda.europa.eu/news-and-events/news/2018/09/13/cyber-ranges-federation-project-reaches-new-milestone

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

Securing 5G Communication in Joint Operations Between NATO Partners

Bruno Dzogovic
Research Scientist / Associate Professor
Telenor ASA / Oslo Metropolitan University
Research & Innovation Department / Department of Computer Science
Oslo, Norway
bruno.dzogovic@telenor.com
bruno.dzogovic@oslomet.no

Silke Holtmanns
Telecommunication Security Expert
Helsinki, Finland
silke@holtmanns.eu

Abstract: NATO considers 5G a "priority area" and the NATO Communication and Information Agency has identified four key areas for the usage of 5G in defence. Currently, each NATO member and defence company has its own approach to using 5G, but it is clear that the defence sector will have to cooperate with public network operators. When using 5G in joint NATO activities, it is important to consider the 5G security approach of each allied partner.

A NATO 5G slice is one promising approach to facilitate cooperation among partners across countries and regions. Commonly, personnel who attend missions in other countries use roaming services. This may expose sensitive and classified information to third parties. Slicing can take place at the application layer, radio access network, core and/or transport level. We will describe the security trade-offs, including roaming and possible improvement approaches, based on the example of a joint NATO operation using 5G slicing. But 5G slicing is only one approach to improving the security of a joint operation. Other approaches include local private networks.

Private networks perform excellently in terms of flexibility, privacy, backhaul usage and reduced network administration. Therefore, military units can use private 5G deployments to connect battlefield units to Command & Control centres and share information among allied parties. This and the various technologies available (e.g., permanent identity protection, legacy usage, shared infrastructure and 5G security

feature usage) have a strong impact on the security and flexibility of the use of 5G in defence.

We will discuss the technology options and their realistic security and practical impacts. Many of those security aspects will be under the control of a public operator, not NATO.

Keywords: *5G security, slicing, joint operations*

1. INTRODUCTION – 5G IN DEFENCE

With the advent of 5G in the industrial and commercial sectors, many verticals have reaped the benefits of advanced connectivity. Be it the Internet of Things, sensors, fixed or mobile broadband, or other entities, industries are witnessing a wide range of applications being introduced into production, commercial operations, or research and development.

The defence sector too is expressing interest in the broad palette of advancements the 5G ecosystem offers. 5G can be customized for specific use scenarios – it offers broadband, low latency, high reliability and support for a large number of connected devices and sensors. With the potential to employ private mobile communications, the military can now deploy non-public 5G independent of mobile network operators or in conjunction and collaboration with them in what the 3rd Generation Partnership Project (3GPP), which defines telecommunications standards, calls a public network integrated – non-public network (PNI-NPN). NATO considers 5G and 6G a "priority area" and an "emerging and disruptive technology" [1].

There is a wide range of potential defence use scenarios, but the NATO Communications and Information (NCI) Agency has identified four key areas [2]:

1) Deployable communications and information systems (CIS) for expeditionary operations
2) Tactical operations
3) Maritime operations
4) Static communications

The NCI is now supporting the NATO Headquarters Consultation, Command and Control (C3) staff [3] and the Allied Command to:

1) formulate a consolidated strategy to create awareness and exercise influence on the civilian-led 5G ecosystem (specifically, influencing 5G standardization) and
2) investigate the benefits and enablers of 5G for military operations as well as to develop and validate concepts for NATO capabilities and drive digitalization in NATO, including the latest developments such as open radio access network (O-RAN).

These strategic considerations now need to be mapped against practical use and deployment.

In Section 2, we will dive into the practical use of 5G in defence, what we can expect in terms of use scenarios and what a joint operation might look like. Section 3 is about mobile technologies, how they work and what security they offer (e.g., 5G slicing). Section 4 tackles how security can be ensured in defence scenarios. Some of these methods are technical, while others are contractual. Section 5 applies the knowledge from Section 4 to the example of a joint operation. Section 6 summarizes this article.

2. PRACTICAL USE OF 5G IN DEFENCE

A. Use Scenarios for 5G in Defence

Defence system manufacturers usually focus on the use of 5G as a communications channel to a local command and control centre, as part of a CIS. But the strength of the NATO alliance lies in its cooperation, which implies that 5G technology is used by different partners, in different countries and in different ways. The NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) studied [4] the different use scenarios and identified potential risks of using 5G to support military movements in a joint operation.

Defence system vendors started creating 5G prototypes for various scenarios. Mobile network-enabled drones are now common [5] and are an important tool for both sides in the Ukrainian war. Here, even the choice of the subscriber identity module (SIM) card becomes a strategic question [6]. A Ukrainian SIM card allows the drone to operate in areas that only have Ukrainian mobile network coverage, thus bypassing some protection mechanisms Ukrainian mobile network operators have put in place [7]. 5G applications are being used in unmanned ground vehicles [8], local networks [9], maritime communication [10], aircraft [11], terrestrial trunked radio replacement [12], and more. Combining mobile 5G networks with satellite communication brings further potential benefits and use scenarios that are being actively researched [13], [14], [3] in the context of 5G advanced and 6G.

Many of the 5G usages in the defence sector are prototypes or testbeds. While they offer important lessons, typically, a proof of concept is created to explore the potential of a certain technology and evaluate its possibilities and usage. Security is rarely on the agenda for a prototype. 5G networks were not designed to meet defence security requirements, but now they are being used for such purposes.

Articles on 5G use in defence scenarios are often accompanied by a note of caution or questions about the resilience and security of the system. After all, 5G is an open standard with application programming interface (API) details published as part of the standards for civil use, not a secret proprietary technology for high-risk scenarios [15]. The defence sector is aware [16], [17] of the general security challenges related to 5G and is working to improve the situation. For example, funding has been provided for a challenge created by the NATO Defence Innovation Accelerator for the North Atlantic [18] to foster innovation and startups to create a security industry that ensures new emerging and disruptive technologies are secure.

B. Practical Considerations in Using 5G in Defence

In the past, security concerns focused on radio jamming and attacks using the interconnection network between mobile operators for spying attacks. The attack scenarios in 5G are now much more complex and diverse due to the evolution of technology that uses virtualization and the opening of public networks to partners. Each 5G defence use case has its own attack and risk profile, depending on the architecture and the nature of the usage. 5G brings many advantages in terms of high-speed, low-latency and secure communication. Nevertheless, these benefits must be balanced against security repercussions, which are of paramount significance.

There is substantial scope to implement 5G in non-public networks. Private 5G can be instituted for public protection, disaster relief and first responders amid catastrophic events that may impair the functioning of society. It can also be used in conflict and war when the military requires important communication resources to support the extensive range of military applications with adequate quality of service (QoS). Such private 5G applications can be deployed tactically, and so are valuable for the defence sector. Today, we already have QoS classes for emergencies. If a person establishes an emergency call and the network is congested, then another person is “kicked out” as the emergency call has higher priority.

One of the key features of 5G is the possibility to use commercial off-the-shelf user equipment (COTS UE) and standardized network functions. A major advantage of COTS UE is its reduced cost, which can substantially reduce overall military expenses in various scenarios. An example of successful COTS UE use in conflict [19] and war settings is in the Russo-Ukrainian war, where mobile communications have proven

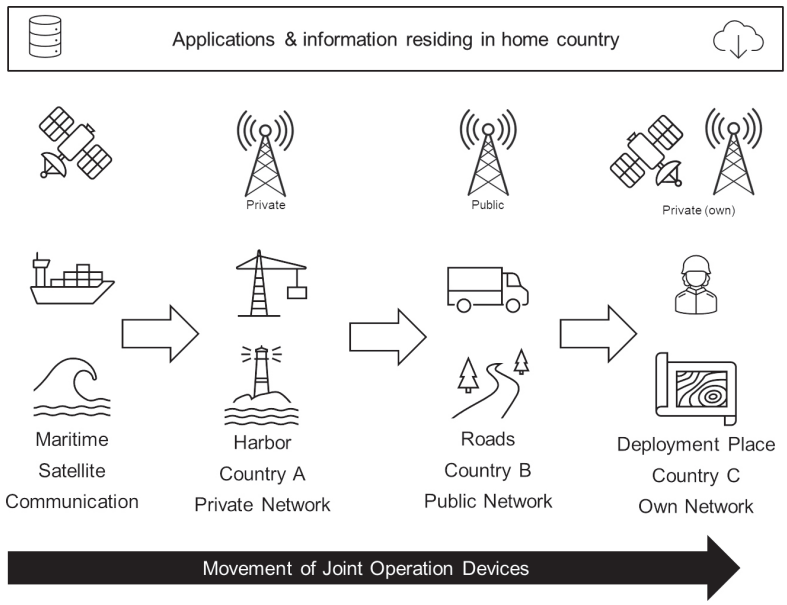
to be an efficient way to collect evidence and intelligence, and exchange tactical information between troops and command centres, providing location services, the ability to stay informed on the battlefield and even performing counter-intelligence operations.

C. Joint Operation Usage Scenario

The scenario we will use to study 5G's security impacts and their solutions is a joint operation strongly aligned with the CCDCOE report [4]. The joint operation has participants from various countries who will bring devices such as unmanned vehicles, phones, and drones with embedded SIMs that belong to different networks. We assume that the devices and the SIM cards are 5G-enabled.

In our scenario, the devices arrive on a ship in country A at a local harbour with a private network and then cross the border into country B by rail or road. The devices reach the site of deployment in country C, where they interact with applications in their home country as well as with partners and their devices from other countries. The partners' devices also potentially use applications in their country of origin (see Figure 1).

FIGURE 1: 5G DEFENCE SCENARIO FOR JOINT OPERATION



3. MOBILE NETWORK TECHNOLOGIES

In our scenario, the device arrives first at the harbour's private network (also called a dedicated or non-public network). Let us assume that it is a 5G network that offers "guest access" to the harbour's customers. After this, the device may temporarily connect to a public network while on the road or rail. That public network could be a long-term evolution (LTE) network (4G), a 5G non-standalone (NSA) network, a 5G standalone (SA) network or a 5G SA network that supports network slicing. The device then crosses the border into country B and switches to a different public network.

We face the following technology challenges in this scenario:

- 4G and 5G interworking network architectures
- SA and NSA networks
- Networks with slicing support and ones with no slicing support
- Private networks and public networks

A. 4G versus 5G

The device may connect to a legacy public network (4G LTE network) when it leaves the harbour. 4G networks have a consumer market-focused security approach. The 4G network itself is considered a security zone whose main security perimeters are the air interface and, to some degree, the interconnection link to other mobile operators. The devices are authenticated, and in most countries, the confidentiality and integrity of the communication between the network and the devices is protected. But 4G does, in some cases, use a permanent international mobile subscriber identity (IMSI) over the air interface, which allows tracking of individual devices. This poses the risk of military equipment movements and potentially targeted strikes being monitored. Also, on the interconnection link, so-called Signalling System No. 7 (SS7) protocol or diameter protocol attacks can be used for location tracking, one-time password interception or data interception [20].

5G has improved security that prevents user tracking on the air interface [21]. While 5G has also improved the security of interconnection between operators, many challenges remain in that area due to intermediaries (interconnection providers, IPX) and the fact that commercial rollout of 5G APIs to operators is still not expected in the near future. Today, user traffic between public mobile networks is not cryptographically protected and passes several IPX providers between the visited mobile network operator and the home mobile network operator, raising questions about confidentiality. The routes of the traffic are usually determined by the cost of data transport. We will assume that

the roaming interface is not properly protected and that various threat vectors will potentially conduct attacks in the future.

B. SA versus NSA Network

Public mobile network operators in our scenario can deploy 5G in SA and NSA modes (i.e., a 4G core with a 5G radio network). In SA mode, that 5G new-radio access is deployed along a fully functional 5G core network, so the communication is considered exclusively 5G. SA mode does not support older devices with 3G and 4G LTE interfaces. It utilizes new types of universal SIM (USIM) cards to support its new security procedures. These new security procedures consist of subscription concealed identifier or subscription permanent identifier (SUCI/SUPI) key pairs that conceal the permanent identity (IMSI) of the user. An SA mode 5G network is required to provide high protection against IMSI catchers and location tracking, and so uses the SUCI/SUPI authentication enhancement.

The mobile operators in the home countries of the joint operation participants in our scenario may issue USIM cards that support 5G SA mode, or they may only issue legacy 4G USIM cards. It is worth noting that the new 5G USIM modules, which support SUCI/SUPI concealment, are backward compatible with 4G LTE networks and NSA modes of operation. Therefore, it is possible that some members of the joint operation are not protected against location tracking using IMSI catchers.

An NSA network comprises 5G base stations that are connected to a 4G or combined 4G/5G core network. An NSA network supports all devices and SIM cards, and operators may decide to gradually roll out 5G using the NSA infrastructure, which has the older authentication procedures and security features. This is done to support a wider range of devices. We expect that many operators support legacy cards through a combined 4G/5G core that lets them serve high-revenue inbound roamers. For the joint operation, it is important to understand what kind of air security the mobile operator of the connected network has and evaluate the risks.

C. Use of Slicing

Slicing is often seen as a solution to isolate sensitive customers inside the network [22]. A slice is a logical and potentially physical division of the network and its resources to provide a specific functionality or service, as in the case of the joint operation. Many parts of the network can be sliced [23]:

- **Device slicing** can be done on the modem, operating system or application level. Currently, we use modem-centric slicing, and this is not expected to change soon. Device slicing isolates information flows inside the device.

- **Transport network slicing** works inside of the mobile operator network. This is currently not common and could potentially be achieved through data network protocols [24], but discussions with progressive operators show that we cannot expect this to be widely supported. Manipulating transport networks dynamically requires substantial automation mechanisms involving software-defined networks (SDNs) and obtaining network intelligence for automatic management, reconfiguration and autoscaling.
- **Radio access network (RAN) slicing** is the most common form of slicing and if slicing is supported by an operator, it is often in the form of RAN slicing to serve specific customer segments with bandwidth- or latency-related QoS requirements. While this gives good availability and latency, and also offers isolation on the radio path of the communication, user communication in the core network is still unencrypted and not isolated between customers.
- **5G core slicing** provides the slice with a dedicated network function, but since most networks still use 4G nodes, 5G core slicing is not so common. To gain most of the benefits of slicing and automation at the core network, 5G infrastructure needs to be deployed in SA mode.
- **Roaming slicing** relies on common attributes as defined in GSM Association (GSMA) specification NG.116 [25] and end-to-end slicing agreements between operators that follow NG.135 [26]. While there is some guidance, we expect there to be many non-standardized variants in the future due to the variety of use cases.

  The different slicing options have different market penetration and security trade-offs as described in Table I.

TABLE I: 5G SLICING OPTIONS AND IMPACTS

	Device slicing	Transport network slicing	RAN slicing	5G core slicing	Roaming slicing
Market	Not available	Not available, but technically possible	Most common type of slicing	Not widely available	Not available commercially
Security trade-offs	Complex implementation in the device	Expensive for operator, especially if no return on investment	Medium complexity	Legacy core elements cannot be used; expensive for operator	Impacts roaming networks all over the world
Security gain	Isolation against other device applications	Isolation against other data flows on transport layer	Isolation on air interface	Isolation against other customers	Isolation independent of network (if properly standardized)

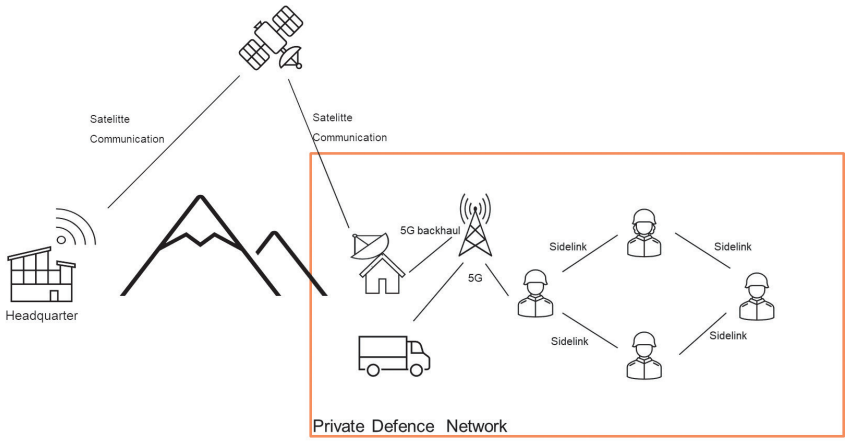
A combination of RAN and core slicing is sometimes called end-to-end slicing, but the term is used loosely. This slicing approach is not risk-free, especially if used in combination with legacy infrastructure [27], [28] or if no end-to-end slicing is considered. In those cases, no real end-to-end security can be guaranteed. In addition, 5G networks rely heavily on cloud and virtualization, which is a new technological leap for the telecommunications industry and poses a potential risk.

D. Private 5G Deployments

Private 5G networks can have different architectures. It can be a dedicated network that is not connected to any public mobile network operators, satellites or even the internet. This kind of dedicated network would be like an isolated bubble. The term dedicated network is also used for private networks that do not connect to any public mobile networks or satellites but do have a data connection to the internet. This form of internet-connected private network is the most common private network today, but it has the drawback that the connection is lost if the user is out of coverage of the private network.

In the joint operation, the devices that arrive at the harbour may use the guest access to the harbour's private network to connect to the internet and applications in their home countries (Figure 1). The joint operation may have its own dedicated network. In this dedicated private network, the ground forces may use 5G sidelink (direct link communication) and satellite communication to headquarters (HQ) (Figure 2).

FIGURE 2: PRIVATE DEFENCE NETWORK AT THE PLACE OF DEPLOYMENT



As an alternative to the satellite link, the private network can also be connected to a public mobile network to offer constant connectivity. This can be achieved either

through a dedicated virtual private network (VPN) or through a roaming connection and the interconnection network (IPX). In the first case, the "donating" public operator could offer roaming via its network, while in the second, the private network would be like any other public network connected to the IPX roaming network. If the private network is connected to the IPX like a normal public network, it can be targeted by SS7 attacks, which are common on IPX. If it is connected via a VPN, attacks may be executed via the "donating" operator.

In general, satellite communications are considered insecure. There are commercial devices that support direct satellite connections, and some operators use satellite communications for the link between the base station and the core (called backhaul) to connect base stations in remote areas. These non-terrestrial networks are an active standardization item in the 3GPP and a discussion of their security would need a separate article.

4. AVAILABLE SECURITY FEATURES AND THEIR USE

While there are many threat angles and risks to consider, there are also many countermeasures that can mitigate them. We will list a range of methods to reduce the risks for joint operations, but there is no silver bullet if existing public networks are used.

A. Protection Against Location Tracking on the Air Interface

Devices should support 5G and the SA mode, which ensures a device only connects to a network that supports the location, privacy and identity protection feature of 5G and safeguards against location tracking and identity theft through IMSI catchers. This prevents the use of disabled location privacy in cases of NSA deployments. There are initial devices [29] that allow explicit monitoring of whether 5G SA is used and restrict communication to the availability of the SA mode. This should be considered for the procurement of SIMs and embedded universal integrated circuit cards for military devices.

B. Protection Against Data Sniffing in the Core Network and Between Networks

Mobile data is secured on the air interface, but after that it is protected at best hop-by-hop between the various network elements. In many cases, such as NSA, 4G and non-5G roaming networks, the mobile data of the user will not in any way be protected between the network elements, which may even belong to different operators and might be routed through the territory of "unfriendly entities". Besides ensuring that the operator partners actually switch on their security features and protocols, application

or transport layer security (i.e., VPN, transport layer security version 1.3) is strongly recommended.

C. Protection Against Jamming and Advanced Interception Attacks

In the long term, we must study whether transmission security (TRANSEC) can be used for mobile network communications at the site of deployment. TRANSEC is a component of communications security and refers to the methods and measures implemented to safeguard communications against interception, cryptoanalysis and in general compromising factors that can help the adversary. The three key components of TRANSEC are:

- Low probability of interception
- Low probability of detection
- Resistance to jamming – electronic protective measures (EPM) and electronic counter countermeasures (ECCM)

By making the system emit lower electromagnetic signatures, it is possible to reduce the probability of communications detection and interception. The communications systems can be targeted by long-range means such as guided missiles, cruise missiles or artillery systems. Therefore, it is of prime importance to reduce the probability of the source of battlefield communications being detected, as the adversary can take advantage of this to intercept communications and engage in decryption or exploit various vulnerabilities in the system. It is possible to use electronic warfare, such as communications jammers, which emit a high-power signal, to jam the communications source and perform denial of service for battlefield communications. For that, the system should be designed with jamming resilience in mind, providing multiple mechanisms to mitigate threats of a similar nature. While TRANSEC provides many advantages, interoperability with public mobile networks and the COTS UE cost advantage might be lost.

D. Mobile Operator Partnering and Selection

Defence manufacturers often partner with mobile network operators [30]. Security should be seen as an integral part of such contracts. The mobile operator should adhere to best security practices. Besides enabling the technical features mentioned in the preceding sections, this could mean vetting the operator according to the 5G Security Control Matrix [31] and the 5G Toolbox [32], auditing its cloud infrastructure, enabling core network internal security, using interconnection firewalls with the latest threat intelligence, pen-testing (core, RAN, cloud, transport, external interfaces, roaming) the network, securing the supply chain, and ensuring a software bill of material is in place and that suppliers (e.g., cloud providers) adhere to highest security standards

and standardized secure software development practices for the entire lifecycle of their products.

The mobile operator should only use certified equipment. Alongside the 3GPP Security Capability Assurance Standards (SCAS) [33], several security certifications and regulations are available or are under development in the European Union that improve the security of 5G networks, such as the EU Cyber Resilience Act, NIS2, EU Common Criteria Certification, EU Cloud Services Certification and EU 5G Certification. They should ensure that the integrity of the subscriber profile is protected and should raise the alarm if sensitive parameters change in a way that might allow attacks (e.g., group memberships, data traffic or SMS/data redirects).

On an operational level, the partnering mobile operator should provide dedicated QoS classes for defence purposes. Potentially, different subcategories can be established depending on the defence situation. For joint operations, the path that the mobile data travels between the visited network and the home network of the device is important. The operator should ensure that the data only travels through friendly nations, ideally through direct connections. SDNs and suitable roaming routes can make this possible.

E. Use of Slicing for a NATO 5G to Extend Coverage and Availability

Private 5G can also be combined to use the public network to extend the connectivity and coverage. On site, the private network would be used, and outside it, the dedicated NATO slice of the public network would be used.

In PNI-NPN, the private component of the network can be controlled, managed and provisioned by the mobile network operator on behalf of the defence owner of the private 5G slice. The responsibility for the management of the PNI-NPN slice can be delegated to the entities involved from the private component of the network and the mobile network operator in parallel, based on a service level agreement (SLA). PNI-NPN mobility with public networks is a relatively new feature (Release 18) in 3GPP and might not be widely available for some time.

F. Use of Slicing Security

The GSMA specifies the "descriptors" of slices in their document NG.116 [26]. These descriptors (generic network slice templates) include attributes such as uplink and downlink bandwidth, as well as aspects such as isolation and 3GPP mission-critical service support. While some aspects are explained sufficiently, others are still slated "for further study", and some have not been considered at all.

For example, isolation can take place at the physical level, transport level, RAN level, core network (user and control plane) level and roaming level. In addition, the

isolation can be logical, through containers, ports and virtual machines, or physical, through different servers and infrastructure. Often when operators mention "slicing", the focus is on RAN slicing only, but that would not offer a sufficiently high degree of isolation for a NATO slice. These aspects are currently not defined and require further work. Aspects such as QoS, priority level and simultaneous use of the network slice attributes can be used to secure a NATO joint operation slice. A NATO slice should at least be logically isolated on the RAN and core network level. Physical isolation is expensive, and while isolation on the transport level is technically possible, general interest in it is currently low.

Security aspects, such as the granularity of OAuth tokens (down to the IMSI and slice level), are currently not part of NG.116. Classes for "legacy" are also not defined, so if a node or intermediate network does not support 5G slicing, it lowers the security level. Other features, such as UE route selection policy to ensure the special handling of defence traffic, would need to be supported by the operator for proper traffic isolation if modem-based device slicing is used.

Partnering public operators should discuss with NATO the required level of granularity and the security aspects, so that the right attributes can be defined in GSMA. General support for 5G protocols on the roaming interface is not expected in the near future. Measures such as end-to-end security between operators (i.e., not hop-by-hop) and "pinning" of routing through friendly nations could be part of a GSMA NG.135 [25] in the future. Currently, roaming routes are determined by factors such as costs and reliability but not security. NATO could consider different requirements for different confidentiality classes.

In their report [34] on network slicing, the US National Security Agency (NSA) and Cybersecurity and Infrastructure Security Agency (CISA) describe key design criteria for network slices that can be combined with the aspects mentioned above to create a secure NATO joint operation slice.

G. Private Network Security Improvements

Mobile networks were not designed for military purposes; therefore, the security standards and processes are potentially not up to the level expected for military use. While there are standards to ensure a baseline degree of security for mobile networks, those standards do not cover all elements of the network and they are indeed only a baseline. Nevertheless, a private network used by the military should conform to the basic product security standards of 3GPP SCAS [33]. Any kind of self-declared security compliance from vendors should be validated either by the defence entities themselves or by an independent third party. The European Union has several good guidelines and documents to improve the security of public mobile networks, which

can be customized and applied to some degree to mobile networks used for defence (e.g., 5G Security Control Matrix [31], 5G Toolbox [32]). For private networks, the same certification considerations apply as for public mobile network equipment.

For mobility and extended coverage, a connection to a public mobile operator network is essential. Using a VPN in a direct link to an operator with good security measures reduces the risk of being attacked via IPX. The security measures of the mobile operator should be validated through compliance audits and an SLA, which should explicitly define the measures the operator must have in place (e.g., signalling firewall with threat intelligence, SIP firewall, cloud security controls, e.g., [35] / C5 [36] or similar compliance). Currently, standards such as C5 have been brought into NATO [37] to protect information. Up-to-date interconnection signalling firewalls are of special importance, as this is a commonly used line of attack today to track persons of interest.

Consequently, the zero-trust model should be considered a long-term goal over the standard perimeter security model. We think of 5G/6G infrastructure as a dynamic, heterogeneous network, and the complexity of such structures renders the perimeter model insufficient and obsolete. The dynamism and scalability of the next-generation networks require more stringent security measures, and thus the zero-trust paradigm becomes an important aspect of security considerations. The US National Security Agency and CISA described further the need for the zero-trust model to provide architectural specifications that introduce additional security layers for deployments that carry confidential traffic, noting that the capabilities and options for a network slice may vary by operator and this method does not address zero trust beyond the slice. A baseline of security-related network slicing features must be established for day-to-day operations. Those features must support confidentiality, integrity and availability requirements. The zero-trust architecture methodology can be implemented to ensure the secure activation, supervision, reporting, modification and de-activation of a slice [34].

5. SECURING THE PATH

Coming back to our example, how can our joint operation be secured? Before the joint operation takes place, we must ensure that the devices have enabled SA mode. A clear policy should be in place that clarifies the communication patterns, security requirements and matching classifications. Here we outline one way to secure the communications. There are many possible variations on this, and as technology and security features advance, better ways of securing the path will become possible.

The devices arrive at the harbour of a friendly country. That harbour has a 4G LTE network with internet access. Some devices are allowed to connect to it for low-security-classification communication. They use a VPN to connect to applications and communicate through the internet access provided by the harbour. Other devices with SA mode enabled note that this is a 4G network, which is prohibited by their policy, and so do not connect to it.

When leaving the harbour, the devices enter the coverage area of a partner public network operator that provides a specific QoS and a dedicated slice in a 5G SA network for the joint operation. This operator also has a sufficient level of security, ensured through an SLA according to the suggestions made above, and has also been audited to ensure the deployment of those security features. The devices connect to this public network slice, and the connections are additionally secured with application and transport layer security.

When arriving at the place of deployment, the ground troops and devices use a private network and direct device-to-device communication – a new radio sidelink or PC5 link (sometimes called direct communication). As far as technically feasible, the devices use TRANSEC. This private network is owned and operated by the joint operation defence team. The private network connects to a communications satellite to link with HQ and has additional security measures as the satellite link is not considered secure.

6. CONCLUSION

A joint operation that uses 5G will face many security challenges. The use of 5G in defence needs to be planned carefully. 5G was designed for civilian use, and the standards and guidelines provide only a limited level of security. But for the defence sector, clear guidance from NATO members on the security expectations and the related use is paramount. The risk involved in each approach needs to be studied and mapped against the NATO security classifications and use. Detailed SLAs with operators and cloud partners need to be created to ensure secure interworking and use of public networks and managed private networks.

Slicing offers some degree of isolation and security for a joint operation, but it requires very specific security support from the mobile network operator hosting the slice. Many of the required security features and slicing attributes are not yet widely available or commonly supported by public mobile networks. The availability of those features depends strongly on market demand and the return on investment. Many mobile operators act when they see a clear market need. This is also true for

the security requirements of the defence sector. Cooperation with mobile operators to work on those features jointly would potentially improve availability.

Other useful security features are still not fully standardized, and the standardization process would benefit from concrete inputs from the defence sector on their requirements to enable the production of standardized, economical COTS UE that can be used in sensitive and high-risk environments. We did not discuss O-RAN in this article, as a proper security discussion of O-RAN in defence would require a separate article due to the complexity of the ecosystem. Many challenges remain for 5G, such as missing standardized features, support from operators, example contracts and security measurement performance indicators, but if 5G for defence wants to use public standards and public networks, those challenges must be addressed.

REFERENCES

[1] "Emerging and disruptive technologies". NATO. Accessed: Jun. 2023. [Online]. Available: https://www.nato.int/cps/en/natohq/topics_184303.htm#policy

[2] NATO Communications and Information (NCI) Agency. "NATO tech Agency explores the potential of 5G for the Alliance". NCIA. Accessed: Jan. 2021. [Online]. Available: https://www.ncia.nato.int/about-us/newsroom/nato-tech-agency-explores-the-potential-of-5g-for-the-alliance.html

[3] G. Capela, "NIC Agency Update on 5G Work", *NITECH NATO Innovation and Technology Journal*, no. 9, pp. 88–89, July 2023. [Online]. Available: https://issuu.com/globalmediapartners/docs/nitech9_-_full_pdf_final?fr=xPf81NTU

[4] V. Oeselg et al., *Research Report Military Movement Risks from 5G Networks*, Tallinn: NATO CCDCOE, 2022. [Online]. Available: https://ccdcoe.org/library/publications/research-report-military-movement-risks-from-5g-networks/

[5] "Lockheed Martin, Verizon demonstrate 5G-powered ISR capabilities for Department of Defense". Lockheed Martin. Accessed: Sep. 2022. [Online]. Available: https://news.lockheedmartin.com/2022-09-28-Lockheed-Martin-Verizon-demonstrate-5G-powered-ISR-Capabilities-for-Department-of-Defense

[6] E. Priezkalns. "Russian attack drone had Ukrainian network SIM for guidance or remote control". Commsrisk. Accessed: Dec. 2023. [Online]. Available: https://commsrisk.com/russian-attack-drone-had-ukrainian-sim-believed-to-have-been-used-for-guidance-and-control/

[7] C. McDaid. "The mobile network battlefield in Ukraine—Part 1". ENEA. Accessed: Mar. 2022. [Online]. Available: https://www.enea.com/insights/the-mobile-network-battlefield-in-ukraine-part-1/

[8] "Protected mobility and defence systems autonomous systems". Patria Group. Accessed: Dec. 2023. [Online]. Available: https://www.patriagroup.com/products-and-services/protected-mobility-and-defence-systems/autonomous-systems

[9] P. Tucker, "The US Navy is testing 5G for future forward operating bases". Defense One. Accessed: Jul. 2022. [Online]. Available: https://www.defenseone.com/technology/2022/07/navy-testing-5g-future-forward-operating-bases/375164/

[10] "LMT, RBF to bring maritime 5G to the Baltic Sea". RCS Wireless News. Accessed: Jun. 2022. [Online]. Available: https://www.rcrwireless.com/20220615/5g/lmt-rbf-to-bring-maritime-5g-to-the-baltic-sea

[11] "Lockheed Martin, AT&T demonstrate 5G high speed transfer of Black Hawk data to 5G.MIL® Pilot Network". AT&T. Accessed: Sep. 2022. [Online]. Available: https://about.att.com/story/2022/5g-lockheed-martin.html

[12] M. Pulliainen, "Virve hyppää 5g-aikaan: Viranomaisten laajakaistaisen mobiiliverkon käyttöönotto alkaa 2022". Tekniikka & Talous. Accessed: Dec. 2021. [Online]. Available: https://www.tekniikkatalous.fi/uutiset/virve-hyppaa-5g-aikaan-viranomaisten-laajakaistaisen-mobiiliverkon-kayttoonotto-alkaa-2022/a3f260b0-0129-40f0-8d45-b284ce0b6951

[13] "Features of 5G from space". Lockheed Martin. Accessed: Dec. 2023. [Online]. Available: https://www.lockheedmartin.com/en-us/products/5g-from-space.html

[14] M. Alleven, “Ericsson, Qualcomm test space-based 5G with Thales”. Fierce Wireless. Accessed: Jul. 2022. [Online]. Available: https://www.fiercewireless.com/tech/ericsson-qualcomm-test-space-based-5g-thales
[15] M. DeGrasse, “5G for defense: U.S. military wants open interfaces, compact infrastructure”. Fierce Wireless. Accessed: Mar. 2023. [Online]. Available: https://www.fiercewireless.com/5g/5g-defense-us-military-wants-open-interfaces-compact-infrastructure
[16] S. Aken. “What does the military’s move to 5G mean for security?” Spiceworks. Accessed: Jul. 2022. [Online]. Available: https://www.spiceworks.com/tech/networking/guest-article/what-does-the-militarys-move-to-5g-mean-for-security/
[17] “Report on national security implications of 5G networks”. US Naval Institute. Accessed: Mar. 2023. [Online]. Available: https://news.usni.org/2023/03/17/report-on-national-security-implications-of-5g-networks-2
[18] “NATO’s innovation accelerator becomes operational and launches first challenges”. NATO. Accessed: Jun. 2023. [Online]. Available: https://www.nato.int/cps/en/natohq/news_215792.htm
[19] K. Freese, “Tradoc: Smart phones playing prominent role in Russia-Ukraine war”. Operational Environment Enterprise. Accessed: Aug. 2023. [Online]. Available: https://oe.tradoc.army.mil/2023/08/10/smart-phones-playing-prominent-role-in-russia-ukraine-war/
[20] P. Donegan, “Threat intelligence in telecoms”. Harden Stance. Accessed: Jul. 2022. [Online]. Available: https://www.hardenstance.com/wp-content/uploads/2022/07/HardenStance-Briefing-Using-Threat-Intelligence-in-Telecoms-2022-FINAL-Subscribers.pdf
[21] P. K. Nakarmi, O. Ohlsson, and P. Hedman. “Fighting IMSI catchers: A look at 5G cellular paging privacy”. Ericsson. Accessed: May 2019. [Online]. Available: https://www.ericsson.com/en/blog/2019/5/fighting-imsi-catchers-5g-cellular-paging-privacy
[22] M. Malik, A. Kothari, and R. A. Pandhare. “Network slicing in 5G: Possible military exclusive slice”. International Conference on the Paradigm Shifts in Communication, Embedded Systems, Machine Learning and Signal Processing (PCEMS), June 2022. Accessed: [Online]. Available: https://ieeexplore.ieee.org/document/9807927
[23] CISA. “5G network slicing: Security consideration for design, deployment and maintenance”. U.S. Department of Defense. Accessed: Jul. 2023. [Online]. Available: https://media.defense.gov/2023/Jul/17/2003260829/-1/-1/0/ESF%205G%20NETWORK%20SLICING-SECURITY%20CONSIDERATIONS%20FOR%20DESIGN,%20DEPLOYMENT,%20AND%20MAINTENANCE_FINAL.PDF
[24] A. Farrel et al. “A framework for network slices in networks built from IETF technologies”. Internet Engineering Task Force (IETF). Accessed: Oct. 2023. [Online]. Available: https://datatracker.ietf.org/doc/draft-ietf-teas-ietf-network-slices/
[25] GSM Association (GSMA), “NG.135 E2E network slicing requirements”, Version 3.0, Jun. 2023.
[26] GSM Association (GSMA), “NG.116 generic network slice template”, Version 8.0, Jan. 2023.
[27] S. Holtmanns and C. McDaid. “5G Network Slicing Security in 5G Core Networks”. ENEA. Accessed: Mar. 2021. [Online]. Available: https://www.enea.com/insights/white-paper-slicing-security-in-5g/
[28] NSA and CISA. “ESF Potential Threats to 5G Network Slicing”. U.S. Department of Defense. Accessed: Dec. 2022. [Online]. Available: https://media.defense.gov/2022/Dec/13/2003132073/-1/-1/0/POTENTIAL%20THREATS%20TO%205G%20NETWORK%20SLICING_508C_FINAL.PDF
[29] “How to turn on 5G standalone mode in iOS 16.4 and enable its advantages”. Mac Observer. Accessed: Feb. 2023. [Online]. Available: https://www.macobserver.com/tips/how-to/turn-on-5g-standalone-mode-enable-advantages/
[30] D. Mortimore. “Launch of secure maritime 5G”. NPS America’s SLAMR. Accessed: Aug. 2022. [Online]. Available: https://nps.edu/web/slamr/-/secure-maritime-5g-ribbon-cutting
[31] “5G security control matrix”. ENISA. Accessed: May 2023. [Online]. Available: https://www.enisa.europa.eu/publications/5g-security-controls-matrix
[32] “5G toolbox”. ENISA. Accessed: Jan. 2020. [Online]. Available: https://www.enisa.europa.eu/topics/critical-information-infrastructures-and-services/telecoms/5g
[33] “Security assurance specifications (SCAS)”. 3GPP. Accessed: Dec. 2023. [Online]. Available: https://portal.3gpp.org/Specifications.aspx?q=1&WiUid=790015
[34] NSA and CISA. “5G network slicing: Security considerations for design, deployment, and maintenance”. CISA. Accessed: Jul. 2023. [Online]. Available: https://www.cisa.gov/news-events/alerts/2023/07/17/nsa-cisa-release-guidance-security-considerations-5g-network-slicing
[35] “Cloud security control matrix”. Cloud Security Alliance (CSA). Accessed: Dec. 2023. [Online]. Available: https://cloudsecurityalliance.org/research/cloud-controls-matrix/

[36] “Cloud computing C5 criteria catalogue”. Federal Office for Information Security (BSI). Accessed: 2020. [Online]. Available: https://www.bsi.bund.de/EN/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Empfehlungen-nach-Angriffszielen/Cloud-Computing/Kriterienkatalog-C5/kriterienkatalog-c5_node.html

[37] S. Niedtfeld, “BSI and NATO: Shaping cloud security in the Alliance”, *BSI Magazine 2022/2: Security in focus*, pp. 45–46, Feb. 2022.

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

Resilience and Vulnerability of Consumer Wireless Devices to Cyber Attacks

Pēteris Paikens
Senior Researcher
Institute of Mathematics and Computer Science
University of Latvia
Riga, Latvia
peteris@ailab.lv

Krišjānis Nesenbergs
Researcher
Cyber-Physical Systems Laboratory
Institute of Electronics and Computer Science
Riga, Latvia
krisjanis.nesenbergs@edi.lv

Abstract: As consumer wireless devices, such as wearables, smartphones and Internet of Things devices become more and more intertwined in our everyday lives, the potential attack surface and the risks if such devices are compromised rise drastically.

Specifically, most of these devices use wireless communication, which uses a broad range of protocols—such as Wi-Fi, Bluetooth and Bluetooth Low Energy—and mesh protocols, such as Zigbee. While Wi-Fi security and vulnerabilities are widely researched and known due to their existing impact on office computing security, the vulnerabilities in Bluetooth and other protocols have received limited attention in the IT security industry because they have historically been hard to monetize for financially motivated threat actors, but these vulnerabilities are still relevant in espionage and cyber conflict. As the prevalence of such devices grows and the costs of equipment such as software-defined radios fall, these vulnerabilities and the related preventive measures need to be better understood.

In this paper we analyze these threats and provide a classification of vulnerabilities and relevant resilience approaches in consumer wireless devices, based on an analysis of the Common Vulnerabilities and Exposures (CVE) reports from 2023 in order to evaluate the risks posed by them to society both in peacetime and during conflicts with a cyber component.

Keywords: *wireless, Internet of Things (IoT), Bluetooth, vulnerabilities, Common Vulnerabilities and Exposures (CVE)*

1. INTRODUCTION*

The classical approach to information system security is primarily concerned with formally structured organizations and technologies [1], so informal activities and human factors are often neglected in practice, even though some related attack vectors, such as social engineering, are highly visible and thus well-known even if not well-understood [2]. This has led to the current situation, where there is a good understanding of the security of systems involving formally structured and well-understood technologies, such as Ethernet and Wi-Fi access points forming a network that is explicitly managed by an organization, while more flexible wireless communications where people are capable of moving devices, easily bringing new devices into the network and acting in less predictable ways with almost limitless potential ad-hoc configurations of devices are often perceived as inherently insecure [3]. Even though such a perception should logically motivate greater scrutiny of wireless communications and their security, the opposite is often true. A lot of myths at both ends of the spectrum—such as "wireless connections are always insecure" or "encrypted wireless connections are secure"—have permeated the common understanding, leading to radical approaches where either all wireless devices are disallowed, or wireless devices are not considered at all in the security threat models. This common lack of understanding usually leads to a wide attack surface, but the relative complexity of exploiting wireless communications has led to a lower number of well-known practical attacks, leading to a false sense of safety that does not withstand scrutiny in the face of more elaborate and capable state actors participating in cyber conflicts.

This has been slowly changing for more mature technologies, such as cell phone protocols [4]–[6] and Wi-Fi [7]–[9], where more control and understanding have been accumulated over the years. Still, the resilience of these technologies is highly reliant on minimizing the attack surface, which is still relatively easy when handling a limited diversity of organization-issued devices with a limited number of known wireless connections. Unfortunately, the advent of smaller, more energy-efficient Internet of Things (IoT) devices and the related differentiation in specialized needs for wireless communications has motivated device manufacturers to adopt new, less mature protocols that lack verified, secure implementation mechanisms.

The problem of IoT device security is even more severe in the consumer market, which has received much less attention in terms of security than devices purchased and maintained in the corporate and government domains. There is a sustained growth in the quantity and variety of cheap wireless consumer IoT devices entering the market and these devices accompany their users almost everywhere—in their homes and workplaces, during transit and in leisure time.

* This research was funded by the Latvian Council of Science project "Automated wireless security analysis of wearable devices" (WearSec), project No. lzp-2020/1-0395.

The global consumer IoT device market in 2023 was evaluated at around $183 billion, with a long-term expected compound annual growth rate (CAGR) of about 5% per year. It is expected to reach $192.4 billion in 2024 and will more than double in less than 15 years [10], making it comparable to the global computer peripherals market, which was worth $470.95 billion in 2022 with an expected CAGR of 6.5% [11]. This is exacerbated by the proliferation and expected rapid growth of wearable smart devices (devices with smart functionality that are worn or carried on the person of the user)—the wearables market was worth $61.30 billion in 2022, with an expected CAGR of 14.6% [12]—due to the increased privacy and surveillance risks of such devices.

Many people currently carry around not only a smartphone but multiple smart embedded or IoT devices with wireless connection capabilities. If a malicious actor takes control of such devices, they can create a variety of security risks—they can serve as Trojan horses into secure infrastructure [13]; become sources of distributed denial-of-service (DDoS) attacks [14], [15]; allow the extraction of secret information leading to industrial espionage [16], political espionage [17] and extortion [18]; and could also potentially hold malware from advanced persistent threats or state actors [19]. There are also the privacy and surveillance risks of tracking or fingerprinting specific devices using their wireless communications [20], [21]. The straightforward solution of just having a policy to remove every wearable consumer device is feasible only in highly controlled environments, and even in this case, there may be issues with devices like medical implants that have similar risks but cannot be removed or sometimes even detected [22]. These risks and the lack of social resilience in case they should be abused on a large scale in cyber warfare motivate the work done in this paper, where we analyze and classify the vulnerabilities and threats of devices to provide a basis for mitigating them in the future.

2. WIRELESS DEVICES AND PROTOCOLS

To explore and classify the threats to consumers, we first have to identify the specific protocols of interest that are used by the consumer wireless devices.

The most widespread wireless protocols in consumer devices are Bluetooth Classic (BT) and Bluetooth Low Energy (BLE), with more than 7 billion devices shipped in 2023 that have one of these protocols enabled [23]. BT has a wide range of applications in audio devices, mobile devices, and certain IoT and Smart Home technologies.

BT is a highly complex set of protocols with support for many different (and contradictory) use cases. Most devices do not need its full capabilities and only

implement some parts of the BT protocol standards. The current certification and testing practices of manufacturers mostly concern reliability in the face of noise or accidental transmission errors, with limited testing for resilience to malicious inputs. In the last few years, multiple critical vulnerabilities in popular BT chipsets have been identified, confirming these gaps in security practices [24]. This situation is made worse by the fact that BLE modules, SoCs (systems on a chip) and devices commonly persist with the same designs for many years without providing upgrade options [25].

A competing proprietary protocol usually used for consumer sportswear is the Adaptive Network Topology (ANT) protocol (and its low-power version, ANT+) by Garmin. This protocol is currently known to be in use in more than 1,000 consumer products, including multiple Samsung mobile phones [26]. ANT is a multicast protocol meant for personal area networks, and thus has some optimizations, such as tree topology, that allow faster, low-energy data rates from wearable sensors than comparable BLE solutions. It uses adaptive isochronous transmission to allow many devices to communicate concurrently without interference, while BLE uses scatternets and broadcasting for the same effect.

Another well-known wireless communication technology available in many consumer devices is radio-frequency identification (RFID) and the family of connected protocols. RFID technology involves tags that are usually passive (although not always) and active readers. Many of the tags can be made read-only, but more and more tags are also "active" in order to improve security through rotating keys and have the option to program them wirelessly. There are multiple RFID technologies based on the frequencies used—low frequency, high frequency (also known as near-field communication or NFC) and ultra-high frequency.

In addition to these mostly well-known technologies, there are several other IoT-related wireless communication protocols mostly made for specific tasks. A group of technologies for long-range IoT communication called low-power wide-area networks includes such protocols as LoRaWAN and Sigfox. LoRaWAN is a point-to-multipoint networking protocol that uses LoRa's physical modulation scheme and hardware [27]. For closer distances or local area networks, Z-wave and Zigbee are frequently used and have interesting security implications [28]. Zigbee is a family of protocols with the standard number IEEE 802.15.4 that is used mostly for home automation and is capable of very low-power communication. Z-wave is also mostly used for home automation, but due to its lower frequency range of about 800–900 MHz, it is capable of much longer-range transmissions and there are more than 4,000 different products in the market that use this protocol. For even closer-range or personal area networking in IoT, one of the most-used technologies is 6LoWPAN [29], which is meant for IPv6 networking over low-power wireless personal area networks and thus can work on

top of IEEE 802.15.4 protocols. Finally, there are several other lesser-known personal area network protocols, some even meant for body area networks, which are all joined under the IEEE 802.15 protocol family.

3. SURVEY OF VULNERABILITIES

The cases listed in the introduction and survey papers [30]–[32] show a variety of threats to organizations and individuals. However, we wanted to contrast them with an analysis of the technical vulnerabilities reported for relevant wireless devices, based on the public MITRE Common Vulnerabilities and Exposures (CVE) list [33].

A. Selection of CVE Reports

In this paper, we review vulnerabilities initially reported throughout 2023, selecting all entries with a 2023 ID as of January 8, 2024,[1] that match specific keywords, analyzing each report to identify relevant aspects and classifying them according to their properties.

The following keywords were used for the initial selection: *IoT, Bluetooth, BLE, BT, ANT, ANT+, LoRa, LoRaWAN, Zigbee, Z-wave, NFC, RFID, 6LoWPAN, IEEE, 802.15, 802.15.1, wireless* and *wearable*. Based on this, a total of 216 vulnerability reports were identified.

Next, these vulnerability reports were analyzed for relevance to the scope of this paper. In a few cases, the keyword search results included unrelated products whose descriptions mentioned the search terms by coincidence or vulnerabilities in software packages that would not be used on the relevant devices but rather on the servers or desktop computers used to manage them. For wireless vulnerabilities, we focused on risks to consumer devices that do not overlap with computers and corporate devices. This excluded Wi-Fi routers, repeaters, access points and Wi-Fi chipset vulnerabilities, because they have been well-studied elsewhere and because of their widespread use in sensitive commercial networks. We excluded vulnerabilities in Windows drivers, but we did consider Linux and Android vulnerabilities as relevant, because those platforms are used not only for computers and smartphones but are also widely used by manufacturers as a basis for many other types of consumer devices. Out of the initial 216 CVE reports we analyzed, 163 were determined to describe vulnerabilities applicable to consumer wireless devices.

B. Analysis and Classification

A limitation of CVE reports is that many of them reflect fixes for bugs with potential

1 As there is a delay between initial CVE report and the public disclosure, at the time of publication will be more vulnerabilities with 2023 IDs, for example, https://www.cve.org/CVERecord?id=CVE-2023-5253 was published at 2024-01-15 after the data collection and analysis.

vulnerabilities that may not be exploitable in real attacks, and none of the reviewed CVEs asserted that these vulnerabilities have been actually exploited "in the wild."

Furthermore, the CVE reports use the Common Vulnerability Scoring System (CVSS) standard to quantify the severity of the vulnerability. While this standard is very useful as a universal qualitative metric, the categories used are designed in the context of "mainstream" vulnerabilities in software running on networked computers, but the physical aspects of wireless protocols and the specifics of consumer wearable and IoT devices require a more targeted approach. For example, CVSS vector "adjacent" (AV:A) is used both for vulnerabilities that require the attacker to simply be in physical range for the wireless connection to function and for vulnerabilities that apply only to previously paired devices, which is a substantial difference with respect to the risk of practical exploitation.

However, the CVE records include links to technical advisories that often are sufficient to manually determine the relevant properties of the vulnerability. Where the technical aspects of the vulnerability were not sufficiently detailed, we made reasonable conservative assumptions to interpret them. Where the CVE did not specify whether a software bug in processing some wireless protocol data could be triggered by a remote attacker or only from the local side, we assume that such data could be delivered remotely.

4. VULNERABILITY CLASSIFICATION

Due to the focus on consumer devices and the potential applications to cyber conflict, we consider it relevant to separate different aspects of classification—the impact, limitations of the attack vector and cause of the problem—instead of reusing existing threat taxonomies.

A. Classification According to Impact

From the perspective of risk analysis, the primary grouping of vulnerabilities is with respect to their potential impact for exploitation and the capabilities that they could offer an attacker, with the relative frequency of these groups shown in Table I.

TABLE I: NUMBER OF CVES ACCORDING TO VULNERABILITY IMPACT

Impact category	Number of CVEs
Information disclosure	53
Denial of service	32
Elevation of privilege	29
Device takeover	42
Unclear	7

1) Information Disclosure

The lowest impact vulnerabilities are those that leak some information that should not be normally accessible. The actual information may vary from a few bytes following some buffer—which might not be useful or dangerous in any way—to capabilities to read arbitrary data from protected system memory that could include encryption keys or other credentials.

2) Denial of Service

Denial-of-service vulnerabilities are limited to temporary disruption of device activities, denying use of devices or disrupting a service. While this does theoretically present a risk, the motivation for potential attackers is limited, as the wireless attacks are limited by range, and only in very niche cases are these devices used in critical scenarios where a temporary disruption of the device would cause significant damage or present a significant gain for the attacker.

Conceptually, IoT devices may have denial-of-service vulnerabilities that allow the attacker to permanently disrupt device operations, which can either require a "factory reset" operation that might not be easily accessible to the operator, restoring the firmware to a known good state or in some cases even "bricking" the device because the chips cannot be restored to normal operation. However, none of the reviewed 2023 CVEs reported a capability for permanent damage.

3) Elevation of Privilege

Many vulnerabilities grant the attacker the capability to do something that they should not have permission to do, such as breaking the operating system user account restrictions model or gaining access to restricted hardware features. Locally exploitable vulnerabilities can present a practical risk in conjunction with another vulnerability that provides arbitrary code execution in a restricted application context. They are also relevant for platforms that enable downloading untrusted third-party applications

or plugins with the expectation that they will be executed within a restricted sandbox environment, but such a vulnerability may enable to break this containment.

4) Device Takeover

The most dangerous group of vulnerabilities are those with the potential to enable the attacker to take control of the device's behavior, either through its own capabilities or by obtaining remote code execution, effectively permitting the takeover of the consumer wireless device. In the context of cyber conflict, this permits the use of these devices for espionage, theft of confidential data and other intelligence operations. We abstain from adopting the widely accepted term "remote code execution" in this context, as a vulnerability may enable attacker-controlled code execution from a different component on the same physical system, not something that is actually remote. Also, there are vulnerabilities that allow the attacker to take over control of the key functions of the device (for example, remotely altering the strength of electrostimulation in a medical device, as in CVE-2023-26979, or opening a smart lock, as in CVE-2023-34625) without necessarily having the ability to execute arbitrary code on the device.

B. Classification According to the Limitations of the Vulnerability

For consumer wireless devices, the two key aspects are the requirement for physical proximity and the requirement for specific conditions (often, the device being paired with the attacking device, which may require user interaction to put the device in pairing mode or approve the connection) for the vulnerability to be exploitable. The riskiest class of vulnerabilities are those that can be exploited remotely over the internet, usually through an exploitable online service, unless the device is directly accessible with a publicly routable IP address. However, in this paper, we focus on vulnerabilities through direct wireless connections, which are grouped as shown in Table II.

TABLE II: NUMBER OF CVES ACCORDING TO ACCESS VECTOR LIMITATIONS

Access vector	Number of CVEs
Remote	63
Remote for a paired device	32
Local	82
Unclear	6

1) Exploitable Locally

Multiple reported vulnerabilities were flaws in the interaction between multiple system components (i.e. the operating system and a Bluetooth controller), so exploitation is possible only if another part of the system is malicious or compromised—this is generally not applicable for remote attackers who do not have physical access, unless combined with another vulnerability. We do not consider risks of backdoored malicious devices or vulnerabilities that require destructive physical access, since in that case the device could be replaced with a malicious equivalent even if the device model does not have any specific vulnerabilities. However, there is also a large group of local vulnerabilities that circumvent the various sandboxing and permission mechanisms on platforms that allow third-party applications to be run (e.g. Android) but expect their capabilities to be limited. Exploitation of those vulnerabilities may allow a seemingly benign application downloaded from an application store to elevate privileges and use the device for malicious purposes, such as surveillance.

2) Exploitable Remotely over the Air in Specific Conditions

Many wireless vulnerabilities require specific conditions that are unlikely to occur in the real world and cannot be easily caused by the attacker. They are still relevant, as they indicate bugs that should be fixed and may become more easily exploitable in conjunction with other vulnerabilities (e.g. a pairing-mode-only vulnerability can be enabled by a different vulnerability that breaks the existing connection, forcing the device to enter pairing mode), but on their own, they do not imply a risk for the device user. However, evaluating this difference for reported CVEs was difficult, as not all security bulletins provided sufficient information about the preconditions to access the vulnerability.

3) Exploitable Remotely over the Air at Any Time

The final and most dangerous class of vulnerabilities are those that can be exploited over the air, using the applicable wireless protocols that require some physical proximity, but are not restricted by the need for the vulnerable device to be in a specific unusual mode or configuration.

C. Classification According to the Cause of Vulnerability

It is also relevant to group vulnerabilities according to what type of problem created it, as that determines the applicable ways to eliminate or at least detect such flaws. The CVEs often (but not always) have some technical information about the nature of the flaw, and during our analysis, we attempted to map these causes to Common Weakness Enumeration (CWE) [34] IDs as maintained by MITRE and group these causes as shown in Table III.

TABLE III: NUMBER OF CVES ACCORDING TO CAUSE OF THE VULNERABILITY

Cause of the vulnerability	Number of CVEs
Memory safety	105
Improper access control	32
Cryptography flaws	14
Unspecified	12

1) Memory Safety

The most common cause of the reviewed vulnerabilities was various types of memory safety issues—buffer overflows and different types of out-of-bounds access. Within this category, we saw CVEs with various issues grouped under CWE 119 (Improper Restriction of Operations within the Bounds of a Memory Buffer), such as:

- CWE 120: Buffer Copy without Checking Size of Input ("Classic Buffer Overflow")
- CWE 125: Out-of-bounds Read
- CWE 126 Buffer Over-read
- CWE 787: Out-of-bounds Write
- CWE 824: Access of Uninitialized Pointer
- CWE 416: Use After Free

We also saw memory issues following CWE 190 (Integer Overflow) or CWE 129 (Improper Validation of Array Index).

All of these are classic software engineering issues that have been largely mitigated in desktop software through decades of investment in tooling, training and engineering policies, but as this data illustrates, they are the currently dominant challenge in consumer device cybersecurity. While it is practically inevitable that not all software is perfect and some bugs will be present, the dominance of these types of issues can be prevented (though at a cost) by the organizations developing the software.

2) Improper Access Control

In this category, we grouped various issues relating to mistakes in verifying the authorization for specific actions or, in some cases, the total lack of any verification. This refers to CWE 284 (Improper Access Control) and its subgroups, such as:

- CWE 862: Missing Authorization
- CWE 306: Missing Authentication for Critical Function
- CWE 648: Incorrect Use of Privileged APIs
- CWE 346: Origin Validation Error
- CWE 20: Improper Input Validation
- CWE 441: Unintended Proxy or Intermediary ("Confused Deputy")

In this category we also included multiple bounds-checking errors if they resulted not in a memory safety issue but triggered a business logic flaw, circumventing some restrictions.

These mistakes are especially relevant when the platform is expected to run untrusted third-party code, such as downloaded apps, and the application programming interface (API) design needs to ensure that security restrictions are enforced for potentially malicious apps.

3) Cryptography Flaws

The final relevant group of vulnerability causes were various flaws relating to the design or implementation of cryptography, or the lack of any cryptography mechanism where one would be reasonably required to prevent the attack. Weaknesses in this group observed in our analysis include:

- CWE 321: Use of Hard-coded Cryptographic Key
- CWE 294: Authentication Bypass by Capture-replay
- CWE 347: Improper Verification of Cryptographic Signature

There was also a set of attacks ("BLUFFS," Bluetooth Forward and Future Secrecy [35]) targeting the cryptographic fundamentals of Bluetooth session encryption keys.

5. RISKS AND RESILIENCE

The vulnerability analysis in the previous section shows that there is an abundance of low-hanging fruit—relatively unsophisticated vulnerabilities caused by well-known risk factors—so attacks are likely not limited by attacker capabilities or the security of the systems but rather by the lack of attacker motivation. A relevant factor affecting motivation is the effect that a successful attacker can hope to achieve, since for many vulnerabilities the impact is limited only to denial of service (19%) or information disclosure (32%). But as 26% of the reported vulnerabilities do show a potential for device takeover, motivation should not be a prohibitive obstacle. Therefore, apparently, the main relevant restriction that leads to the low level of observed attacks is the

requirement for physical proximity, which makes it hard to perform mass attacks. This limitation makes these vulnerabilities relevant only to attackers who intend to target a specific person or a limited number of people located relatively close to the attacker. The proximity requirement also acts as a deterrent by making it clear to a potential attacker that they might be identified and penalized for any malicious acts.

A. Threat Model

With that in mind, the main threat model relevant for these vulnerabilities of consumer wireless devices to cyber attacks is a sophisticated attacker, possibly a state-sponsored actor, who intends to attack existing vulnerable consumer devices in the target country with the goal of either disrupting civilian life or specifically targeted espionage.

The other threat model is targeted attacks for personal reasons, especially within the context of domestic disputes and violence, which is an established motivation for the abuse of technology [36], and one where the attacker's goals explicitly include surveillance and control of IoT devices, rather than fraud or other forms of monetization.

It is important to note that the threats in this model are largely speculative, because while the risks and vulnerabilities are there, the motivation for such attacks is limited, as there are not many options to exploit them for financial gain or perform them at a large scale due to the physical proximity requirement. The IoT exploits in the wild mentioned in the reports of major security vendors are limited to compromised IoT devices becoming part of botnets and being used to attack other systems that attackers consider valuable via DDoS [37], [38], once again demonstrating the industry focus on protecting corporate networks and commercial services.

B. Resilience

The list of reported vulnerabilities in consumer wireless devices is dominated by memory safety issues—buffer overflows, out-of-bounds access, use after free—even more than in the case of desktop software. However, the development practices applied to embedded systems seem to lag behind other domains of software development, and the situation will likely improve with diligent application of the same best practices: a thorough review of static analysis tools and compiler features to identify potential risks in C or C++ source code, and gradual switch in from C/C++ to memory-safe systems programming languages such as Rust or Golang. Also, fuzzing is a powerful approach to discovering implementation flaws and vulnerabilities, as demonstrated by projects such as Frankenstein [39], BrakTooth [24] and SweynTooth [40], and it can be used in integration testing to identify deviations and undocumented features [41] in parts from third-party vendors. Some of the CVE reports analyzed in this paper noted that the issues had been discovered in this manner.

Of course, that will only be applied by the device manufacturers if they have sufficient motivation to do so. For consumers and society in general, resilience relies on measures such as third-party penetration testing during procurement of devices with potentially risky applications, or a liability shift toward making device manufacturers financially responsible for consequences of security flaws, which may motivate them to invest in measures to reduce vulnerabilities.

6. CONCLUSION

We observe that the published vulnerability data overrepresents issues in general-purpose computer systems, as opposed to non-computer devices whose installed base is far larger. We also observe that most of the reported vulnerabilities are for platforms or software development kits, but not for specific devices or products.

To us, the fact that relatively few registered CVE records apply to consumer or IoT devices is not reassuring. Given the relatively large number of relevant wireless security flaws identified in major software platform projects such as Android, Linux kernel and Zephyr project, and the relatively low level of investment in and attention to security of non-computer consumer devices, we would expect that the multitude of custom proprietary systems would also have a comparable or higher number of flaws. However, the lack of reported CVEs indicates that for most IoT products and companies making them, vulnerabilities are either unidentified or identified outside of public view, and any devices are likely to be vulnerable without the general public knowing.

Similarly, the issues reported with a specific software platform would apply to many different products using that platform. However, there is often no simple way to identify specific devices that use that version of the platform and may be vulnerable. Therefore, the reports are useful for device manufacturers if they properly track their software dependencies, but not for the general public protecting itself. Some manufacturers[2] report affected *chipsets*, but it is not easy for consumers to identify which chipsets are used in their devices and whether they are affected, leading to inaction due to the inability to determine which threats are applicable to specific devices.

The events of 2023 have once more demonstrated the interest of advanced actors in achieving surveillance and spyware goals using highly sophisticated malware such as TriangleDB [42] or Pegasus [43] that exploit multiple iOS zero-day vulnerabilities and were detected only years after their first attacks. Due to substantial investment by Apple and Google, it is technically much more difficult to exploit smartphones than

2 For example, Qualcomm, https://docs.qualcomm.com/product/publicresources/securitybulletin/december-2023-bulletin.html.

various consumer wireless devices, in which the CVEs we reviewed often represented the low-hanging fruit of basic vulnerabilities.

While we are happy to see that the detected vulnerabilities were fixed proactively before they were exploited in the wild, this does provoke an important question: Have there really been no sophisticated attacks on these devices, or are we just not able to detect them?

REFERENCES

[1] I. V. Koskosas and N. Asimopoulos, "Information system security goals," *International Journal of Advanced Science and Technology*, vol. 27, pp. 15–26, 2011.

[2] X. Luo, R. Brody, A. Seazzu, and S. Burd, "Social engineering: The neglected human factor for information security management," *Information Resources Management Journal (IRMJ)*, vol. 24, no. 3, pp. 1–8, 2011.

[3] M. B. Schmidt, "Development and analysis of a model for assessing perceived security threats and characteristics of innovating for wireless networks," Ph.D. dissertation, Mississippi State University, USA, 2006.

[4] M. A. Ferrag, L. Maglaras, A. Argyriou, D. Kosmanos, and H. Janicke, "Security for 4G and 5G cellular networks: A survey of existing authentication and privacy-preserving schemes", *Journal of Network and Computer Applications*, vol. 101, pp. 55–82, 2018.

[5] H.-M. Wang, T.-X. Zheng, J. Yuan, D. Towsley, and M. H. Lee, "Physical layer security in heterogeneous cellular networks," *IEEE Transactions on Communications*, vol. 64, no. 3, pp. 1204–1219, 2016.

[6] R. Odarchenko, V. Gnatyuk, S. Gnatyuk, and A. Abakumova, "Security key indicators assessment for modern cellular networks," in *2018 IEEE First International Conference on System Analysis & Intelligent Computing (SAIC)*.

[7] H. Peng, "WIFI network information security analysis research," in *2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet)*.

[8] G. Sagers, B. Hosack, R. Rowley, D. Twitchell, and R. Nagaraj, "Where's the security in WiFi? An argument for industry awareness," in *2015 48th Hawaii International Conference on System Sciences*.

[9] M. Hooper, Y. Tian, R. Zhou, B. Cao, A. P. Lauf, L. Watkins, W. H. Robinson and W. Alexis, "Securing commercial WiFi-based UAVs from common security attacks," in *MILCOM 2016 – 2016 IEEE Military Communications Conference*.

[10] Grand View Research. "Consumer IoT – Worldwide." Statista. Accessed: Jan. 5, 2024. [Online]. Available: https://www.statista.com/outlook/tmo/internet-of-things/consumer-iot/worldwide

[11] "Global computer peripherals market size by devices (input devices, output devices), by connectivity (wired, wireless), by end user (commercial, residential), by geographic scope and forecast." Verified Market Research. Accessed: Jan. 5, 2024. [Online]. Available: https://www.verifiedmarketresearch.com/product/computer-peripherals-market/

[12] "Wearable technology market size, share & trends analysis report by product (head & eyewear, wristwear), by application (consumer electronics, healthcare), by region (Asia Pacific, Europe), and segment forecasts, 2023–2030." Grand View Research. Accessed: Jan. 5, 2024. [Online]. Available: https://www.grandviewresearch.com/industry-analysis/wearable-technology-market

[13] O. Arias, J. Wurm, K. Hoang, and Y. Jin, "Privacy and security in internet of things and wearable devices," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 1, no. 2, pp. 99–109, 2015.

[14] R. Khader and D. Eleyan, "Survey of DoS/DDoS attacks in IoT," *Sustainable Engineering and Innovation*, vol. 3, no. 1, pp. 23–28, 2021.

[15] K. Doshi, Y. Yilmaz, and S. Uludag, "Timely detection and mitigation of stealthy DDoS attacks via IoT networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2164–2176, 2021.

[16] O. D'Mello, M. Gelin, F. B. Khelil, R. E. Surek, and H. Chi, "Wearable IoT security and privacy: A review from technology and policy perspective," in *Future Network Systems and Security: 4th International Conference (FNSS)*, Paris, 2018.

[17] D. Carstens, J. Mahlman, J. Miller, and M. Shaffer, “Mobile device espionage,” Association for Industry, Engineering and Management Systems (AIEMS), 2019.
[18] J. Ibarra, H. Jahankhani, and S. Kendzierskyj, “Cyber-physical attacks and the value of healthcare data: Facing an era of cyber extortion and organised crime,” *Blockchain and Clinical Trial: Securing Patient Data*, pp. 115–137, 2019.
[19] F. Blow, Y.-H. Hu, and M. Hoppa, “A study on vulnerabilities and threats to wearable devices,” *Journal of the Colloquium for Information Systems Security Education*, vol. 7, no. 1, 2020.
[20] Q. Xu, R. Zheng, W. Saad, and Z. Han, “Device fingerprinting in wireless networks: Challenges and opportunities,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 94–104, 2015.
[21] B. Blumbergs, Ē. Dobelis, P. Paikens, K. Nesenbergs, K. Solovjovs, and A. Rušiņš, “WearSec: Towards automated security evaluation of wireless wearable devices,” in *Nordic Conference on Secure IT Systems*, Reykjavik, 2022.
[22] Y. Kim, W. Lee, A. Raghunathan, V. Raghunathan, and N. K. Jha, “Reliability and security of implantable and wearable medical devices,” in *Implantable Biomedical Microsystems*, S. Bhunia, S. J. A. Majerus, and M. Sawan, Eds., Elsevier, 2015, pp. 167–199.
[23] M. Powell. “2023 Bluetooth Market Update.” Bluetooth. Accessed: Jan. 2, 2024. [Online]. Available: https://www.bluetooth.com/2023-market-update/
[24] M. E. Garbelini, V. Bedi, S. Chattopadhyay, S. Sun, and E. Kurniawan, “BrakTooth: Causing havoc on Bluetooth Link Manager via directed fuzzing,” in *31st USENIX Security Symposium (USENIX Security 22)*, Boston, 2022.
[25] M. Cäsar, T. Pawelke, J. Steffan, and G. Terhorst, “A survey on Bluetooth Low Energy security and privacy,” *Computer Networks*, vol. 205, 2022.
[26] Garmin Canada Inc. “ANT+ Directory.” Ant. Accessed: Feb. 12, 2024. [Online]. Available: https://www.thisisant.com/directory/filter/2316/~/~/~/
[27] M. A. Ertürk, M. A. Aydın, M. T. Büyükakkaşlar, and H. Evirgen, “A survey on LoRaWAN architecture, protocol and technologies,” *Future Internet*, vol. 11, no. 10, 2019.
[28] C. W. Badenhop, S. R. Graham, B. W. Ramsey, B. E. Mullins, and L. O. Mailloux, “The Z-Wave routing protocol and its security implications,” *Computers & Security*, vol. 68, pp. 112–129, 2017.
[29] G. Mulligan, “The 6LoWPAN architecture,” in *Proceedings of the 4th Workshop on Embedded Networked Sensors*, 2007.
[30] A. Barua, M. A. Al Alamin, M. S. Hossain, and E. Hossain, “Security and privacy threats for Bluetooth Low Energy in IoT and wearable devices: A comprehensive survey,” *IEEE Open Journal of the Communications Society*, vol. 3, pp. 251–281, 2022.
[31] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, “A survey on Internet of Things: Architecture, enabling technologies, security and privacy, and applications,” *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1125–1142, 2017.
[32] V. Hassija, V. Chamola, V. Saxena, D. Jain, P. Goyal, and B. Sikdar, “A survey on IoT security: Application areas, security threats, and solution architectures,” *IEEE Access*, vol. 7, 2019.
[33] “CVE database search.” MITRE. Accessed: Jan. 2, 2024. [Online]. Available: https://cve.mitre.org/
[34] “Common Weakness Enumeration (CWE).” MITRE. 2023. Accessed: Jan. 5, 2024. [Online]. Available: https://cwe.mitre.org/index.html
[35] D. Antonioli, “BLUFFS: Bluetooth forward and future secrecy attacks and defenses,” in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, Copenhagen, 2023.
[36] J. Slupska and L. M. Tanczer, “Threat modeling intimate partner violence: Tech abuse as a cybersecurity challenge in the Internet of Things,” in *The Emerald International Handbook of Technology-Facilitated Violence and Abuse*, J. Bailey, A. Flynn, and N. Henry, Eds., Emerald Publishing Limited, 2021, pp. 663–688.
[37] E. Altares, J. Salvio, and R. Tay. “FortiGuard Labs threat research – 2022 IoT threat review.” Fortinet. Accessed: Jan. 5, 2024. [Online]. Available: https://www.fortinet.com/blog/threat-research/2022-iot-threat-review
[38] Check Point Research. “The tipping point: Exploring the surge in IoT cyberattacks globally.” Check Point. Apr. 11, 2023. Accessed: Jan. 8, 2024. [Online]. Available: https://blog.checkpoint.com/security/the-tipping-point-exploring-the-surge-in-iot-cyberattacks-plaguing-the-education-sector/
[39] J. Ruge, J. Classen, F. Gringoli, and M. Hollick, “Frankenstein: Advanced wireless fuzzing to exploit new Bluetooth escalation targets,” in *29th USENIX Security Symposium (USENIX Security 20)*, 2020.
[40] M. E. Garbelini, C. Wang, S. Chattopadhyay, S. Sumei, and E. Kurniawan, “SweynTooth: Unleashing mayhem over Bluetooth Low Energy,” in *2020 USENIX Annual Technical Conference (USENIX ATC 20)*, 2020.

[41] J. Classen and M. Hollick, “Inside job: Diagnosing Bluetooth lower layers using off-the-shelf devices,” in *Proceedings of the 12th Conference on Security and Privacy in Wireless and Mobile Networks*, 2019.
[42] I. Kuznetsov, V. Pashkov, L. Bezvershenko, and G. Kucherin. “Operation Triangulation: iOS devices targeted with previously unknown malware.” Kaspersky. Jun. 1, 2023. Accessed: Jan. 5, 2024. [Online]. Available: https://securelist.com/trng-2023/
[43] K. Hawkinson. “Apple warns Russian journalists their phones may be targets of a state-sponsored attack.” Business Insider. Sep. 16, 2023. Accessed: Jan. 5, 2024. [Online]. Available: https://www.businessinsider.com/apple-russia-journalists-pegasus-spyware-putin-ukraine-war-2023-9

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

Identifying Obstacles of PQC Migration in E-Estonia

Jelizaveta Vakarjuk
Cybernetica AS and
Department of Software Science
Tallinn University of Technology
Tallinn, Estonia
jelizaveta.vakarjuk@cyber.ee

Nikita Snetkov
Cybernetica AS and
Department of Software Science
Tallinn University of Technology
Tallinn, Estonia
nikita.snetkov@cyber.ee

Peeter Laud
Cybernetica AS
Tartu, Estonia
peeter.laud@cyber.ee

Abstract: With the development of quantum technologies, there is an urgent need to secure existing IT infrastructure against quantum threats. Introducing post-quantum cryptography (PQC) to existing systems may protect them against future quantum computer attacks. Still, post-quantum migration is a cumbersome process that requires systematic planning and takes years. In this paper, we study Estonia's e-government ecosystem, outline systems and products that rely on potentially vulnerable cryptographic primitives, identify the main migration obstacles, and provide recommendations on how the migration process should be carried out.

Keywords: *post-quantum cryptography, e-governance, migration*

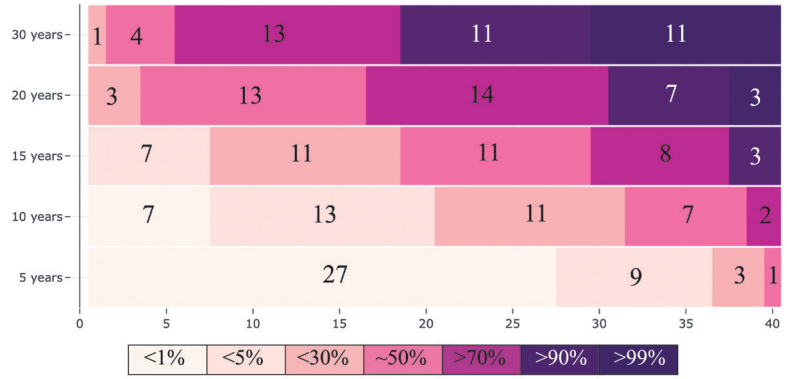
1. INTRODUCTION

In 1994, Peter Shor showed that sufficiently powerful quantum computers can solve integer factorization and discrete logarithm problems whose hardness is the foundation of many modern public-key cryptosystems. Therefore, we have to reckon with the emergence of cryptographically significant quantum computers (CSQCs) [1]. Such computers can eliminate the practical usage of most public-key primitives, such as (EC)DH,[1] RSA,[2] (EC)DSA,[3] and EdDSA[4] [2], and affect the key and/or output sizes of symmetric key schemes such as AES[5] and SHA[6] [3]. For that reason, several standardization agencies and industrial entities initiated the process of migration to post-quantum (also known as quantum-safe) cryptography [4].

One cannot reliably predict a date when a CSQC will be available. Several factors influence progress in this area. The first one is when the *circuit for Shor's algorithm* is optimized. There are different ways in which the quantum circuit for Shor's algorithm can be built [5]–[7]; some require fewer qubits,[7] while others require fewer operations or fewer specific gates. There may be further optimizations of circuits that can influence how soon breaking RSA with keys of practical length becomes feasible. Another is progress in *error correction*, since realizing physical qubits is a non-trivial task. As physical qubits interact with each other, errors appear. Correcting these errors is hard due to the non-cloning theorem [8]. Therefore, the number of physical qubits needed to implement Shor's algorithm is much higher than the number of logical (error-corrected) qubits [9]. This is currently an active research area; reducing the ratio between the logical and physical qubits is an important goal. Finally, *chip architecture* is progressing. There are different types of qubits (e.g., ion traps, photonics, and superconducting qubits) [10], each requiring a different architecture when assembled into chips. Moreover, this architecture may be different even for the same type. All these aspects show that judging the progress of quantum technology development by just the number of announced qubits is not accurate. Instead, leading experts can be surveyed to determine their opinions on how long it will take before a CSQC is built. Such surveys have already been carried out; we have cited the results of one of them [1] in Figure 1.

1 Elliptic Curve Diffie-Hellman (ECDH).
2 Rivest-Shamir-Adleman (RSA).
3 Elliptic Curve Digital Signature Algorithm (ECDSA).
4 Edwards-curve Digital Signature Algorithm (EdDSA).
5 Advanced Encryption Standard (AES).
6 Secure Hash Algorithm (SHA).
7 Qubit is a basic unit of quantum information.

FIGURE 1: OPINION OF 40 EXPERTS ON THE LIKELIHOOD OF HAVING A QUANTUM COMPUTER ABLE TO FACTORIZE A 2048-BIT NUMBER IN 24 HOURS



Estonia is known for its success in e-government, including citizen ID cards, interoperability services, i-voting, and e-taxes. A significant amount of Estonian infrastructure relies on the security of used cryptographic primitives. The process of migrating these services to quantum-safe cryptographic schemes and protocols is a non-trivial task, because post-quantum algorithms have properties different from those of the algorithms currently used. For example, there is no drop-in replacement for the Diffie-Hellman key exchange, or for RSA, that can be used as both a digital signature and an encryption algorithm. The key sizes, signature, and ciphertext sizes are increased, complicating their use with constrained devices such as smart cards.

Related work. Kampanakis et al. [11] identify research gaps and possible standard updates that are required for the PQC migration process. The work focuses mostly on the impact of PQC on authentication in transport protocols and proposes sixteen open questions for the research. Attema et al. [12] created a handbook for the PQC migration process to help different organizations to organize and plan the PQC migration process. It explains concrete steps in the process and gives advice on how to mitigate the threat of quantum computers to their systems. Additionally, there is the Open Quantum Safe (OQS) project [13], which supports the post-quantum migration process by helping with the implementation and evaluation aspects of PQC. OQS maintains a library for the post-quantum cryptographic algorithms, as well as their integration into various protocols and applications, such as OpenSSL.

In this work, we discuss the main obstacles to migrating cryptography-reliant e-Estonia technologies to PQC. We outline the current status of research in post-quantum cryptography and provide recommendations for software/security architects and decision-makers.

2. PRELIMINARIES

A. Priorities in Transitioning to PQC

One of the main goals of e-government services is to assure the confidentiality, integrity, and authenticity of the information exchanged between the governmental institutions, citizens, and businesses. A service provides these properties by relying on various cryptographic primitives (e.g., encryption for confidentiality, signatures for integrity and non-repudiation, etc.). For interoperability between services and their clients, the use of cryptography has to be sufficiently standardized, such that all stakeholders understand the relevant data structures and encodings in the same way.

A breakthrough in the cryptanalysis of contemporary cryptographic primitives, achieved by, for instance, a CSQC, affects whether a system achieves all the security goals mentioned above, but it affects them in quite different ways. The used authentication protocols must be updated before a CSQC is available, but currently, we can continue using existing protocols, because authentication happens in the moment. The evidentiary value of a signature on a digital document can be preserved if someone takes the necessary steps to show that the signature was created before a CSQC came into being. The secrecy of a message encrypted today may be breached if the adversary stores the ciphertext and manages to obtain a CSQC and to recover the plaintext while the obligation of confidentiality remains valid. While the goal of transitioning to PQC is to make sure that a system continues to provide its security properties, this analysis shows that different priorities may be assigned to different properties and to subsystems ensuring these properties.

B. Quantum Key Distribution and PQC

Quantum key distribution (QKD) is a technology that enables parties to establish a shared secret key for exchanging encrypted data [14]. QKD is based on the laws of quantum physics, implying that information exchanged over quantum channels cannot be copied. Any interference in the communication will be noticeable by protocol participants, since the to-be-transferred quantum state is destroyed. Hence QKD is affected by denial-of-service attacks. QKD requires the creation and management of specific and costly infrastructure. For longer QKD networks, several trusted intermediate nodes are necessary. In 2023, the European Quantum Flagship initiated the EuroQCI project to construct quantum communication infrastructure within the European Union.[8] Estonia participates in EuroQCI through the sub-project EstQCI, led by the Ministry of Economic Affairs and Communication.[9]

PQC, however, helps to solve a wider range of problems, offering key establishment, encryption, digital signatures, and so on. Deploying PQC algorithms and testing their performance in the currently used protocols is much easier than QKD, because

8 https://petrus-euroqci.eu
9 https://www.riks.ee/kvantside/estqci-kvantside-projekt

they run on classical hardware. Therefore, experimenting with and deploying PQC is currently more urgent than building QKD networks.

C. NIST Standardization

The National Institute of Standards and Technology (NIST) initiated a standardization competition for post-quantum algorithms in 2016 and received submissions of twenty-three signature schemes and fifty-nine key establishment mechanisms (KEM) built on a variety of mathematical problems.[10] The main families of post-quantum algorithms are lattice-based, code-based, isogeny-based, hash-based, multivariate-based, and based on MPC-in-the-head.[11]

After the third round of NIST standardization competition, seven finalist schemes and eight alternate schemes were selected [15]. Four schemes were selected to become future standards: Crystals-Kyber [16] for the KEM category and Crystals-Dilithium, Sphincs+, and Falcon for the signature category.[12] Crystals-Kyber, Crystals-Dilithium, and Falcon are lattice-based schemes, while Sphincs+ is hash-based. Crystals-Dilithium [17] is considered the primary signature scheme, suitable for all use cases. Sphincs+ [18] is more conservative security-wise but the least efficient. Falcon [19] has the smallest key and signature sizes but requires floating point arithmetic.

Since most of the selected schemes rely on structured lattices, the NIST decided to continue the standardization competition to find alternative schemes. The candidates for the KEM category were selected from the schemes in the fourth round of competition—Classic McEliece [20], BIKE[13] [21], and HQC[14] [22] (all of them code-based). The isogeny-based candidate SIKE[15] was broken. For signature schemes, the NIST announced a new call, receiving fifty submissions.[16] The NIST expects two candidates at most to be selected for the standardization. No candidate is expected to replace Crystals-Dilithium as the primary signature scheme.

The NIST also initiated a separate process for standardizing *stateful* hash-based signatures: Leighton-Micali Signature (LMS) and eXtended Merkle Signature Scheme (XMSS) [23]. Both LMS and XMSS are considered to be secure against quantum computers, but they are less practical than Sphincs+, Falcon, or Crystals-Dilithium. Their main limitation is that the signer must keep track of a state. Protecting the state and backing it up along with the private key is still an open question. One potential solution could be threshold cryptography [24].

10 https://csrc.nist.gov/News/2016/Public-Key-Post-Quantum-Cryptographic-Algorithms
11 Multi-party computation in the head (MPC-in-the-head) is a paradigm that allows to create digital signature that is a non-interactive zero-knowledge proof of knowledge of the secret key.
12 https://csrc.nist.gov/Projects/post-quantum-cryptography/selected-algorithms-2022
13 Bit Flipping Key Encapsulation (BIKE).
14 Hamming Quasi-Cyclic (HQC).
15 Supersingular Isogeny Key Encapsulation (SIKE).
16 https://csrc.nist.gov/csrc/media/Projects/pqc-dig-sig/documents/call-for-proposals-dig-sig-sept-2022.pdf

D. European Standardization and Security Agencies

European organizations like BSI,[17] ANSSI,[18] ETSI,[19] ENISA,[20] and NCSC[21] have also published reports on the transition to post-quantum cryptography, listing algorithms they recommend using and explaining how they should be used. Some of the recommended algorithms are different from the recommendations of the NIST. Table I indicates which algorithms are recommended by which organization.

TABLE I: AGENCY RECOMMENDATIONS

Organization	KEM	Signatures
NIST	Crystals-Kyber	Crystals-Dilithium, Falcon, Sphincs+, XMSS, LMS
BSI [25]	FrodoKEM, Classic McEliece, Crystals-Kyber*	LMS/HSS, XMSS/XMSS MT, Crystals-Dilithium,* Sphincs+*
ANSSI [26]	Crystals-Kyber, FrodoKEM	Crystals-Dilithium, Falcon, XMSS, LMS, Sphincs+
NCSC [27]	Crystals-Kyber	Crystals-Dilithium, Falcon, Sphincs+, XMSS, LMS

* After NIST standards are available

FrodoKEM [28] is a lattice-based KEM that was submitted to the NIST competition but, due to its performance, was not selected. FrodoKEM and Classic McEliece are recommended due to their more conservative and well-understood security. However, both schemes are less efficient than Crystals-Kyber and may not suit all the applications. Additionally, post-quantum cryptography is recommended for use only in *hybrid mode*. Only hash-based signature schemes may be used as standalone solutions. Managing the state of XMSS or LMS is an important concern; it must not be copied or backed up to the other device, because this may lead to a forked state, potentially resulting in security breaches.

E. Hybrid Schemes

In the context of PQC, hybrid mode refers to the usage of post-quantum algorithms together with classical algorithms. Hybrid mode is used to guarantee security even if one of the algorithms gets broken or if an implementation vulnerability is found.

Using a KEM in hybrid mode is theoretically straightforward; one would use a KEM *combiner* that takes as input both ECC[22]/RSA key material and PQC key

17 German Federal Office for Information Security (BSI).
18 French Cybersecurity Agency (ANSSI).
19 European Telecommunications Standards Institute (ETSI).
20 European Union Agency for Cybersecurity (ENISA).
21 National Cyber Security Centre (NCSC).
22 Elliptic Curve Cryptography (ECC).

material and outputs a symmetric key that is computed from both key materials. BSI recommendations for KEM combiners are CatKDF[23] and CasKDF[24] [29] and the NIST's Keccak (SHA3, KMAC[25]) and HMAC[26]-based KDFs [30].

Combining PQC with pre-quantum cryptography in public key certificates is more complicated. Multiple variants have been proposed [31], all with their own limitations (Table II). The most straightforward solution is to use *multiple certificates*, that is, having separate certificates with post-quantum keys and with pre-quantum keys. With this setup, all entities (CA, subCA, client) have two distinct key pairs for the same identity. This solution makes it possible to keep the existing infrastructure and supplement it with a mirror copy based on post-quantum algorithms.

Another option is to use the *AltPublicKey* extension [32] of X.509 certificates, adding a post-quantum key and the corresponding signature. This approach can be used with legacy systems, such that the main signature on the certificate is pre-quantum and verifiable by any device, and the alternate signature may be verified by the parties supporting PQC.

The *chameleon* [33] approach makes it possible to hide one certificate inside another and extract the inner certificate when needed. With this approach, the system can decide whether both signatures should be verified or just one of them.

The *composite* [34] approach makes it possible to define key and signature objects, each of which internally consists of two keys and signatures. This approach allows for adopting post-quantum schemes without changing the logic of application when it is used but instead by changing the cryptographic library that specifies these composite objects and operations with them. The specification [34] was designed to consider composite algorithms to be FIPS[27]-approved even when one of the component algorithms is not. When choosing an appropriate hybrid mode, it is important to understand the system requirements and limitations.

23 Concatenate Key Derivation Function (CatKDF).
24 Cascade Key Derivation Function (CasKDF).
25 Keccak Message Authentication Code (KMAC).
26 Hash-based Message Authentication Code (HMAC).
27 US Federal Information Processing Standard (FIPS).

TABLE II: HYBRID APPROACHES

Approach	Advantages	Disadvantages
Multi-certificate	• No changes to the existing infrastructure (a copy is created). • Can choose when to transmit large post-quantum certificates and signatures.	• Difficult to use with protocols or architectures supporting a single signature or certificate. • Difficult to manage layers.
AltPublicKey	• Compatible with legacy systems. • Compatible with applications that are limited to a single certificate.	• Large keys for post-quantum primitives need to be transmitted even if not used. • Requires updating protocols to verify/ produce multiple signatures.
Chameleon	Large post-quantum keys can be dropped if not used.	Requires updating protocols to verify/ produce multiple signatures.
Composite	• Both keys are used at the same time, offering the best security. • Satisfies regulatory requirements.	Not compatible with legacy systems.

F. Migration to PQC

Migration from classical cryptography to PQC is a resource- and time-consuming process, with a timeline that might exceed five years [12]. Therefore, the migration process should begin as soon as possible. The migration framework introduced in [35] and used later in [12] consists of three main stages:

1) compilation of cryptographic inventory;
2) preparation of the migration plan;
3) execution of the migration plan.

The first stage consists of identifying all locations where cryptographic technologies are being used, including, but not limited to:

1) confidentiality and integrity of data at rest or in transit;
2) authentication of users or other system elements;
3) access control to resources of the system [35].

One must identify what data should be protected and for how long. This makes it possible to determine the urgency of PQC migration and the priorities of migrating different systems. The questions in Annex A.1 of [35] can help in compiling a cryptographic inventory.

In the second stage, the main challenge is to choose which post-quantum schemes should be implemented and how. Not all post-quantum algorithms are suitable for all use cases; one must choose suitable algorithms based on the systems' limitations, constraints, and requirements. Implementing PQC may also require new hardware that supports those algorithms.

In the third stage, the migration plan from the previous stage is executed. In this step, it is crucial to avoid introducing new vulnerabilities during the implementation. Attention should be paid to side-channel resistance of the implemented schemes [36]–[39]. Additionally, it is important to maintain cryptographic agility, which allows for switching between different post-quantum algorithms.

3. CRYPTOGRAPHIC INVENTORY

Many of the services underlying the infrastructure of e-Estonia rely heavily on different cryptographic algorithms; some of them even go beyond regular digital signatures and encryption. Migrating all those services to post-quantum cryptography while preserving interoperability is a non-trivial task, given the challenges of PQC. We start by identifying the systems of e-Estonia that rely on cryptography and the parts of them that a CSQC would break.

We see that for many applications listed in Table III, data privacy needs to be ensured for a long time. These applications may be targets of *harvest attacks*, where the adversary collects encrypted data now and decrypts it later, when quantum computers become available. It may already be too late to prevent harvest attacks, since PQC is not used in current protocols, and adversaries could already be collecting the traffic. Still, the impact of those attacks can be mitigated.

For some digital signature use cases (e.g., signing long-term contracts), the forgery protection must be long-term. Once the adversary is able to forge a user's signature, the authenticity of data protected by this signature is questionable and one has to be careful when accepting signatures under this key pair. We know what it takes to extend the validity period of signatures. Some of it is reflected in the current AdES formats [40] for archival signatures; to achieve the rest, the entity interested in preserving the evidentiary value refreshes the time stamps [41].

TABLE III: USAGE OF CRYPTOGRAPHIC PRIMITIVES WITHIN ESTONIAN INFRASTRUCTURE

Cryptographic scheme	Function	Post-quantum security	Applications in e-Estonia
RSA	Encryption, signature	Broken	Smart-ID, ID card, X-Road
ElGamal	Encryption	Broken	I-voting
ECDSA	Signature	Broken	ID card, Mobile-ID
ECDH	Key establishment	Broken	TLS
AES	Encryption	Key size increase required [42], [43]	TLS, ID card

4. TRANSITION TO QUANTUM-SAFE ALTERNATIVES

The migration process to post-quantum cryptography is more challenging and resource-consuming than the previous cryptographic migrations (e.g., DES to AES, SHA1 to SHA2, RSA to ECDSA after the Estonian ID card crisis[28]). Unfortunately, PQC has no drop-in replacement for ECDH or RSA. There is no post-quantum scheme that offers both encryption and signing functionality as RSA. No scheme with properties similar to the Diffie-Hellman key exchange was submitted to the NIST standardization competition.

As the key and signature sizes of algorithms grow, protection against side-channel attacks increases in importance. Therefore, each application should be handled separately, and an appropriate quantum-safe alternative should be chosen on the basis of its requirements and limitations. For use cases that rely on multiple cryptographic primitives or use non-standard techniques like threshold cryptography for Smart-ID, the transition to quantum-safe primitives is more challenging and time-consuming. In this section, we will identify the main challenges of migrating services to post-quantum cryptography and suggest which post-quantum algorithms are most suitable.

The following challenges and propositions are grouped according to the technologies they apply to, where the technologies we focused on are the most fundamental ones for Estonian e-governance. Indeed, Lips et al. [44], referencing UN e-government surveys [45], [46], identify X-Road as the backbone of Estonian e-government. On top of it, a large number of diverse services have been built in both the public and the private sector. These services use the identification methods provided by X-Road to

28 https://news.err.ee/616732/potential-security-risk-could-affect-750-000-estonian-id-cards

interact with each other, while the end users depend on e-ID and underlying PKI to access them. To this mix we add another significant application: internet voting.

A. Smart-ID

Smart-ID provides users with authentication and digital signing functionality. These are both achieved using the RSA multi-prime signature scheme [47], which has separate key pairs for authentication and signing. Smart-ID protocol relies on threshold cryptography [48], meaning that the private (signing) key is split into two shares: one is stored on the user's mobile device, and the other is stored on the server. To create a signature, the mobile device and the server interact to apply their shares of the private key, producing a single signature that can be verified using a single public key. The main goal of the solution is to offer protection for the private key: an adversary obtaining only one private key share cannot create valid signatures.

The current protocol is built around the RSA signature scheme, because its mathematical structure supports the creation of such protocols. Unfortunately, the structure of post-quantum signature schemes does not allow such protocols to be created easily. Out of the three (future standard) signature schemes, Crystals-Dilithium has the best mathematical structure but includes a few challenging parts. First, it has the rejection sampling step, which aims to verify that the final signature does not leak information about the private key. If the verification does not pass, signing is restarted and repeated until a valid signature is created. The number of restarts is three or four (on average) for the to-be-standardized parameters. When the signing process is split into two parts, both the mobile device and the server must perform rejection sampling, increasing the number of restarts. Second, due to a more complicated signing algorithm, the number of communication rounds needed to produce a signature will be increased (compared to RSA). Vakarjuk et al. [49] attempt to create an alternative to the Smart-ID protocol using a lattice-based signature scheme similar to Crystals-Dilithium. However, unlike the current Smart-ID protocol, the verification algorithm is not the same as that of the standardized scheme.

For authentication, one does not necessarily have to use a standardized cryptographic algorithm. Hence a signature scheme with suitable properties may be chosen more freely. But for signing, compliance with standards is a strict requirement. Therefore, a threshold signature protocol should produce signatures that are verified using the verification algorithm in the standard.

Another approach toward quantum-safe Smart-ID is to use a "threshold-friendly" signature scheme. The NIST may standardize such a scheme in the future [50],[29] but waiting would delay post-quantum migration.

29 https://csrc.nist.gov/csrc/media/Projects/pqc-dig-sig/documents/call-for-proposals-dig-sig-sept-2022.pdf

B. ID Card

The ID card is a state-issued identity document that allows using different e-services. The ID card is a compulsory document for Estonian citizens and European Union citizens who reside permanently in Estonia. The ID card gives its users different functionalities: providing authentication to various services, creating qualified electronic signatures, and encrypting/decrypting documents. The ID card contains two key pairs with corresponding certificates: one for digital signatures and the other for authentication and decryption. ID cards have limited memory and computational power, making the running of PQC algorithms difficult. Table IV presents a key and signature size comparison of pre-quantum and post-quantum algorithms that provide approximately the same level (approx. 128-bit) of security.

TABLE IV: SIZES IN BYTES

Algorithm	Public key	Private key	Signature
RSA3072	400	384	384
ECDSA P-256	32	32	64
Dilithium2	1312	2528	2420
Falcon-512	897	1281	666
XMSS-SHA2_16_256	64	2093	2692
Sphincs+	32	64	7856

For smartcards, protection against side-channel attacks is crucial. However, adding protection against side-channel attacks to the post-quantum cryptographic schemes adds complexity to the algorithms and increases the amount of random-access memory (RAM) needed to execute operations.

If migration via the hybrid approach is chosen, then ID cards would need to support the creation of both post-quantum and pre-quantum signatures, storing all the keys and certificates. Not all solutions from Section 2.E are suitable for smart cards. Memory limits the use of a multi-certificate solution, as it would require storing four certificates on a card. AltPublicKey or chameleon solutions are more suitable, as both allow interoperability with legacy systems while also permitting the creation of post-quantum signatures for updated systems.

The same considerations apply to Mobile-ID, as it also relies on a tamper-resistant chip to protect the key material. Furthermore, reducing the size of the signatures is important, because the communication with servers is SMS-based.

Moreover, the Estonian ID card uses the same key pair for both authentication and decryption [51]. Since we do not have an RSA post-quantum drop-in replacement, we would have to introduce an additional key pair. This leads to one more certificate being stored on the ID card.

C. X-Road

X-Road provides secure data exchange between different information systems in the public and private sectors. The identity of each organization is verified using certificates issued by the certification authorities. Data exchanged using X-Road is protected both at rest and at transit. Since X-Road is used to exchange data between the public sector information systems, long-term data protection is necessary. To hinder harvesting attacks, it is essential to start protecting data using the key derived with a post-quantum key establishment algorithm as soon as possible. BSI, ANSSI, and ETSI recommend using the Crystals-Kyber scheme in hybrid mode with ECDH to provide security against both classical and quantum adversaries.

A main component of the X-Road infrastructure is a *security server* that manages service calls and responses between different information systems. Each security server holds an authentication key pair to establish secure communication channels with other security servers and a signing key pair to sign all outgoing messages. Choosing the right hybrid mode for signing is less straightforward than for the key establishment. Signing and verification should be fast and the signature should be short, due to how signing is used in X-Road. A straightforward way is the concatenation of a pre-quantum (RSA or ECDSA) and a post-quantum (Crystals-Dilithium) signature. Using concatenation to combine two signatures guarantees unforgeability if at least one of the signature schemes is unforgeable [52]. This approach requires modifying security servers to produce and verify two signatures instead of one.

D. Public Key Infrastructure

For PKI, choosing a suitable post-quantum algorithm for digital signatures on the certificates is a challenging task. The hybrid modes for certificates are outlined in Section 2.E. There is also a *mixed architecture* solution that can be considered for the certificate chains. In mixed architecture, algorithms with stronger security guarantees are chosen for the long-lived objects such as root CAs; more efficient algorithms are selected for short-lived objects such as end-entity certificates or TLS handshakes. For example, hash-based signatures like Sphincs+ or XMSS/LMS can be used for root CAs, since they rely only on the security of underlying hash functions. For the other

certificates, schemes like Crystals-Dilithium or Falcon providing smaller signatures can be used. This type of solution would require services to support all the mentioned signature schemes.

The main obstacle Estonia faces in transferring to quantum-safe PKI is that it must rely on the other parties who contribute to the change—hardware security module vendors, certificate authorities, policymakers, and browser vendors.

E. I-voting

In the Estonian internet voting protocol, asymmetric cryptography is used to encrypt and sign the votes [53]. Further cryptographic techniques—mix-nets [54]—are used to break the visible links between individual votes that were cast and those that were counted. In this setting, the signature mechanisms are largely independent of the other used cryptographic constructions, while vote encryption and mix-nets are tightly coupled.

The signatures for votes are generated using the signature creation devices described above and obtain their legal meaning through the public-key infrastructure also described above. Hence, no adaptations specific to i-voting are necessary. The situation is quite different for encryption. Currently, the votes are encrypted using ElGamal encryption, and the mix-net protocol in use [55] has been designed to mix them. Neither the encryption nor the mix-net are post-quantum secure. The migration to PQC primarily involves the introduction of a post-quantum mix-net, which will fix the encryption algorithm that it can support. Constructions exist for such mix-nets, but they either do not have sufficient performance [56] or impose a significant change on the format of the votes and the design of the whole voting protocol [57].

The lack of suitable protocols becomes even more debilitating when considering hybrid approaches. We would need an encryption scheme whose security can be derived either from a well-studied pre-quantum hardness assumption or from a post-quantum hardness assumption. While such schemes can be constructed compositionally, the accompanying mix-nets probably cannot. We are also not aware of any research toward mix-nets for hybrid encryption schemes. Using a hybrid encryption scheme with a mix-net that is able to mix only a single layer of encryption defeats the purpose of using that scheme. At a minimum, it would leak the links between cast and counted votes if/when one of the encryption layers becomes insecure.

5. COMMONALITIES AND DIFFERENCES

Only a few obstacles are common to all the analyzed systems, which have different architecture, security, and regulatory requirements. Changes influencing all systems and applications are the increased size of keys, signatures, and ciphertexts. Additionally, our analysis shows that there is no one scheme that suits all use cases; thus, appropriate quantum-safe alternatives must be chosen based on the requirements and constraints of each individual system.

The following obstacles were identified in this paper:

1) The urgency of starting the post-quantum migration is not well understood by decision-makers and those outside the cryptographic community. Multiple parties from the private and public sectors are not contributing enough to the migration process, causing stagnation.
2) EuroQCI focuses attention on QKD technology, whose functionality is more limited than that provided by PQC.
3) Standardized post-quantum schemes are computationally more complex and storage-heavy and therefore less compatible with smart cards.
4) Side-channel attack protection for the post-quantum schemes is underdeveloped.
5) PQC in hybrid mode limits which hybrid certificate solutions can be deployed on smart cards.
6) The absence of an RSA-like scheme providing both signing and encryption requires changing decryption functionality on ID cards, increasing the code footprint.
7) Choosing suitable hybrid modes for signature schemes is more challenging, as certain security guarantees need to be ensured.
8) Some unexpected obstacles to implementing post-quantum schemes become obvious only in the later stages of the migration process—the implementation and testing phases.
9) Research is lacking on post-quantum cryptography for esoteric use cases such as i-voting and distributed signing (Smart-ID).

6. NEXT STEPS

In Section 4, we have identified the most suitable post-quantum schemes that can be used to replace the currently used cryptography. This can be used to prepare a full migration plan for Estonian e-services, also taking into account services not analyzed in this paper. The proposed post-quantum alternatives (including different

hybrid modes) should be tested within the systems to identify further challenges that are not obvious from the primary analysis. If necessary, other post-quantum schemes can be implemented to verify whether they fit better under the limitations identified during the testing phase. We propose that the (soon-to-be) standardized schemes be considered, because their security has been studied more carefully by the cryptographic community. The migration plan should also outline the order in which the various services are transitioned to PQC. The services dealing with more sensitive data that must stay secret for a long time should be the first to switch to quantum-safe alternatives.

In Section 2.A, we indicated that the migration of confidentiality mechanisms to PQC was the most urgent. Fortunately, it is also the easiest, at least from the point of view of coordination among the stakeholders (i.e., standardization). Indeed, to protect the data at rest, the party holding it may select the mechanisms alone. To protect data in transit, the two parties must agree on the algorithms and formats. They also must have a mechanism to authenticate each other.

Authentication is similar, requiring only the client and the relying party to coordinate. Indeed, proprietary solutions for authentication (e.g., Mobile-ID or Smart-ID) are proliferating even today. Digital signatures for non-repudiation, however, are very different. Here any solution must be compliant with legislation, which requires standardization, certification, and so on.

ACKNOWLEDGMENTS

This research has been supported by the Estonian Research Council, Grant No. PRG1780, and by the European Union under Grant Agreement No. 101087529. The views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

[1] M. P. Michele Mosca. “Quantum threat timeline report 2022.” Global Risk Institute. Accessed: Mar. 11, 2024. [Online]. Available: https://globalriskinstitute.org/publication/2022-quantum-threat-timeline-report/

[2] P. W. Shor, “Algorithms for quantum computation: Discrete logarithms and factoring,” in *Proceedings 35th Annual Symposium on Foundations of Computer Science*, IEEE, 1994, pp. 124–134.

[3] L. K. Grover, “A fast quantum mechanical algorithm for database search,” in *Proceedings of the 28th Annual ACM Symposium on Theory of Computing*, 1996, pp. 212–219.

[4] NIST Computer Security Division. “Announcing request for nominations for public-key post-quantum cryptographic algorithms.” NIST. Dec. 20, 2016. [Online]. Available: https://csrc.nist.gov/News/2016/Public-Key-Post-Quantum-Cryptographic-Algorithms
[5] S. Beauregard, “Circuit for Shor’s algorithm using 2n+3 qubits,” *Quantum Info. Comput*., vol. 3, no. 2, pp. 175–185, Mar. 2003.
[6] U. Skosana and M. Tame, “Demonstration of Shor’s factoring algorithm for N=21 on IBM quantum processors,” *Scientific Reports*, vol. 11, p. 16599, Jan. 2023.
[7] C. Gidney and M. Ekerå, “How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits,” *Quantum*, vol. 5, p. 433, 2021, doi: 10.22331/Q-2021-04-15-433.
[8] W. K. Wootters and W. H. Zurek, “A single quantum cannot be cloned,” *Nature*, vol. 299, no. 5886, pp. 802–803, Oct. 1982, doi: 10.1038/299802a0.
[9] I. Georgescu, “25 years of quantum error correction,” *Nature Reviews Physics*, vol. 2, no. 10, p. 519, Oct. 2020, doi: 10.1038/s42254-020-0244-y.
[10] J. Dargan, “What types of quantum computers exist in 2023?” *Quantum Insider*, Jun. 2023. [Online]. Available: https://thequantuminsider.com/2023/06/06/types-of-quantum-computers/
[11] P. Kampanakis and T. Lepoint, “Vision paper: Do we need to change some things? Open questions posed by the upcoming post-quantum migration to existing standards and deployments,” in *Security Standardization Research – 8th International Conference*, SSR 2023, Lyon, France, Apr. 22–23, 2023, pp. 78–102, doi: 10.1007/978-3-031-30731-7_4.
[12] “The PQC Migration Handbook. Guidelines for Migrating to post-quantum cryptography,” Dutch Organization for Applied Scientific Research, Dec. 2023. [Online]. Available: https://www.tno.nl/en/newsroom/2023/04-0/pqc-migration-handbook/
[13] D. Stebila and M. Mosca, “Post-quantum key exchange for the Internet and the Open Quantum Safe project,” in *Selected Areas in Cryptography (SAC) 2016*, vol. 10532, R. Avanzi and H. Heys, Eds., Springer, 2017, pp. 1–24. [Online]. Available: https://openquantumsafe.org
[14] C. H. Bennett and G. Brassard, “Quantum cryptography: Public key distribution and coin tossing,” *Theoretical Computer Science*, vol. 560, pp. 7–11, 2014.
[15] G. Alagic et al., “Status report on the third round of the NIST post-quantum cryptography standardization process,” U.S. Department of Commerce, National Institute of Standards and Technology, Jul. 2022, doi: 10.6028/NIST.IR.8413-upd1.
[16] “Module-lattice-based key-encapsulation mechanism standard (initial public draft).” FIPS PUB 203. [Online]. Available: https://csrc.nist.gov/pubs/fips/203/ipd, Aug. 2023.
[17] “Module-lattice-based digital signature standard (initial public draft).” FIPS PUB 204. Aug. 2023. [Online]. Available: https://csrc.nist.gov/pubs/fips/204/ipd
[18] “Stateless hash-based digital signature standard (initial public draft).” FIPS PUB 205. Aug. 2023. [Online]. Available: https://csrc.nist.gov/pubs/fips/205/ipd
[19] P.-A. Fouque et al. “Falcon: Fast-Fourier lattice-based compact signatures over NTRU. Specification v1.2.” Falcon. Oct. 2020. [Online]. Available: https://falcon-sign.info
[20] D. J. Bernstein et al. “Classic McEliece: conservative code-based cryptography: Cryptosystem specification.” Classic McEliece. Oct. 2022. [Online]. Available: https://classic.mceliece.org
[21] N. Aragon et al. “BIKE – Bit Flipping Key Encapsulation. Round 4 submission,” Bike. Oct. 2022. [Online]. Available: https://bikesuite.org/
[22] C. Aguilar Melchor et al. “Hamming quasi-cyclic (HQC). Fourth round submission.” PQC HQC. Apr. 2023. [Online]. Available: https://pqc-hqc.org
[23] “Stateful hash-based signatures.” NIST. Accessed: Mar. 11, 2024. [Online]. Available: https://csrc.nist.gov/projects/stateful-hash-based-signatures
[24] J. Kelsey, S. Lucks, and N. Lang, “Coalition and threshold hash-based signatures,” Cryptology ePrint Archive, Paper 2022/241, 2022. [Online]. Available: https://eprint.iacr.org/2022/241
[25] BSI, “Cryptographic mechanisms: Recommendations and key lengths,” BSI Technical Guideline TR-02102-1, Jan. 2023. [Online]. Available: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/TechGuidelines/TG02102/BSI-TR-02102-1.html
[26] ANSSI. “ANSSI views on the Post-Quantum Cryptography transition (2023 follow up).” French Cybersecurity Agency (ANSSI). Oct. 2023. [Online]. Available: https://cyber.gouv.fr/en/publications/follow-position-paper-post-quantum-cryptography
[27] John H. “Migrating to post-quantum cryptography.” National Cyber Security Centre blog post. Nov. 2023. [Online]. Available: https://www.ncsc.gov.uk/blog-post/migrating-to-post-quantum-cryptography-pqc
[28] E. Alkim et al. “FrodoKEM: Learning with errors key encapsulation. preliminary standardization proposal.” FrodoKEM. Mar. 2023. [Online]. Available: https://frodokem.org

[29] “ETSI TS 103 744. CYBER; Quantum-safe hybrid key exchanges.” Dec. 2020. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/103700_103799/103744/01.01.01_60/ts_103744v010101p.pdf
[30] E. Barker, L. Chen, and R. Davis, “Recommendation for key-derivation methods in key-establishment schemes,” National Institute of Standards and Technology, Aug. 2020, doi: 10.6028/nist.sp.800-56cr2.
[31] M. Ounsworth, “Post-quantum multi-key mechanisms for PKIX-like protocols: Problem statement and overview of solution space,” Internet Engineering Task Force, Internet-Draft draft-pq-pkix-problem-statement-01, Sep. 2019. [Online]. Available: https://datatracker.ietf.org/doc/draft-pq-pkix-problem-statement/01/
[32] “X.509: Information technology – Open Systems Interconnection – The Directory: Public-key and attribute certificate frameworks.” International Telecommunication Union (ITU). Oct. 2019. [Online]. Available: https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=X.509
[33] C. Bonnell, J. Gray, D. Hook, T. Okubo, and M. Ounsworth, “A mechanism for encoding differences in paired certificates,” Internet Engineering Task Force, Internet-Draft draft-bonnell-lamps-chameleon-certs-03, Jan. 2024. [Online]. Available: https://datatracker.ietf.org/doc/draft-bonnell-lamps-chameleon-certs/03/
[34] M. Ounsworth, J. Gray, M. Pala, and J. Klaußner, “Composite signatures for use in Internet PKI,” Internet Engineering Task Force, Internet-Draft draft-ounsworth-pq-composite-sigs-11, Dec. 2023. [Online]. Available: https://datatracker.ietf.org/doc/draft-ounsworth-pq-composite-sigs/11/
[35] “CYBER; Migration strategies and recommendations to Quantum Safe schemes,” ETSI, 2020. [Online]. Available: https://www.etsi.org/deliver/etsi_tr/103600_103699/103619/01.01.01_60/tr_103619v010101p.pdf
[36] H. M. Steffen, G. Land, L. J. Kogelheide, and T. Güneysu, “Breaking and protecting the crystal: Side-channel analysis of dilithium in hardware,” in *Post-Quantum Cryptography – 14th International Workshop, PQCrypto 2023*, College Park, MD, USA, August 16–18, 2023, pp. 688–711, doi: 10.1007/978-3-031-40003-2_25.
[37] S. Marzougui, V. Ulitzsch, M. Tibouchi, and J.-P. Seifert, “Profiling Side-Channel Attacks on Dilithium: A Small Bit-Fiddling Leak Breaks It All,” Cryptology ePrint Archive, Paper 2022/106, 2022. [Online]. Available: https://eprint.iacr.org/2022/106
[38] Y. Ji, R. Wang, K. Ngo, E. Dubrova, and L. Backlund, “A side-channel attack on a hardware implementation of CRYSTALS-Kyber,” in *IEEE European Test Symposium*, ETS 2023, Venice, Italy, May 22–26, 2023, pp. 1–5, doi: 10.1109/ETS56758.2023.10174000.
[39] A. Wagner, V. Wesselkamp, F. Oberhansl, M. Schink, and E. Strieder, “Faulting Winternitz one-time signatures to forge LMS, XMSS, or SPHINCS+ signatures,” in *Post-Quantum Cryptography – 14th International Workshop, PQCrypto 2023*, College Park, MD, USA, Aug. 16–18, 2023, pp. 658–687, doi: 10.1007/978-3-031-40003-2_24.
[40] “ETSI EN 319 132-1. Electronic signatures and infrastructures (ESI); XAdES digital signatures; Part 1: Building blocks and XAdES baseline signatures.” ETSI. Feb. 2022. [Online]. Available: https://www.etsi.org/deliver/etsi_en/319100_319199/31913201/01.02.01_60/en_31913201v010201p.pdf
[41] M. Geihs, “Long-term protection of integrity and confidentiality—security foundations and system constructions,” PhD thesis, Darmstadt University of Technology, Germany, 2018. [Online]. Available: http://tuprints.ulb.tu-darmstadt.de/8094/
[42] L. K. Grover, “A fast quantum mechanical algorithm for database search,” in *Proceedings of the 28th Annual ACM Symposium on the Theory of Computing*, G. L. Miller, Ed., Philadelphia, Pennsylvania, USA, May 22–24, 1996, pp. 212–219, doi: 10.1145/237814.237866.
[43] L. Chen et al., “Report on post-quantum cryptography,” National Institute of Standards and Technology Internal Report 8105,” Apr. 2016, doi: 10.6028/NIST.IR.8105
[44] S. Lips, V. Tsap, N. Bharosa, R. Krimmer, T. Tammet, and D. Draheim, “Management of national eID infrastructure as a state-critical asset and public-private partnership: Learning from the case of Estonia,” *Inf. Syst. Frontiers*, vol. 25, no. 6, pp. 2439–2456, 2023, doi: 10.1007/S10796-022-10363-5.
[45] UN Department of Economic and Social Affairs, “United Nations e-government survey 2018 – Gearing e-government to support transformation towards sustainable and resilient societies,” United Nations, New York, 2018.
[46] UN Department of Economic and Social Affairs, “E-government survey – Digital government in the decade of action for sustainable development,” United Nations, New York, 2020.
[47] A. Buldas, A. Kalu, P. Laud, and M. Oruaas, “Server-supported RSA signatures for mobile devices,” in *Computer security – ESORICS 2017*, Cham: Springer International Publishing, 2017, pp. 315–333.
[48] Y. Desmedt, “Society and group oriented cryptography: A new concept,” in *Advances in Cryptology—CRYPTO ’87*, Berlin, Heidelberg: Springer, 1988, pp. 120–127.

[49] J. Vakarjuk, N. Snetkov, and J. Willemson, "DiLizium: A two-party lattice-based signature scheme," *Entropy*, vol. 23, no. 8, p. 989, 2021.
[50] R. Peralta and L. T. A. N. Brandão, "NIST first call for multi-party threshold schemes," NIST – National Institute of Standards and Technology, Jan. 2023, doi: 10.6028/nist.ir.8214c.ipd.
[51] "CDOC 2.0 spetsifikatsioon. v0.9." Cybernetica AS. Jan. 2023. [Online]. Available: https://installer.id.ee/media/cdoc/cdoc_2_0_spetsifikatsioon_d-19-12_v1.9.pdf
[52] N. Bindel, U. Herath, M. McKague, and D. Stebila, "Transitioning to a quantum-resistant public key infrastructure," Cryptology ePrint Archive, Paper 2017/460, 2017. [Online]. Available: https://eprint.iacr.org/2017/460
[53] S. Heiberg, T. Martens, P. Vinkel, and J. Willemson, "Improving the verifiability of the Estonian internet voting scheme," in *Electronic Voting – First International Joint Conference, e-Vote-ID 2016*, Bregenz, Austria, Oct. 18–21, 2016, pp. 92–107, doi: 10.1007/978-3-319-52240-1_6.
[54] M. Abe, "Universally verifiable mix-net with verification work independent of the number of mix-servers," in *Advances in Cryptology – EUROCRYPT '98, International Conference on the Theory and Application of Cryptographic Techniques*, Espoo, Finland, May 31 – Jun. 4, 1998, pp. 437–447, doi: 10.1007/BFB0054144.
[55] D. Wikström, "A sender verifiable mix-net and a new proof of a shuffle," in *Advances in Cryptology – ASIACRYPT 2005, 11th International Conference on the Theory and Application of Cryptology and Information Security*, Chennai, India, Dec. 4–8, 2005, pp. 273–292, doi: 10.1007/11593447_15.
[56] V. Farzaliyev, J. Willemson, and J. K. Kaasik, "Improved lattice-based mix-nets for electronic voting," *IET Inf. Secur.*, vol. 17, no. 1, pp. 18–34, 2023, doi: 10.1049/ISE2.12089.
[57] X. Boyen, T. Haines, and J. Müller, "A verifiable and practical lattice-based decryption mix net with external auditing," in *Computer Security – ESORICS 2020: Proceedings of the 25th European Symposium on Research in Computer Security, Part II*, Guildford, UK, Sep. 14–18, 2020, pp. 336–356, doi: 10.1007/978-3-030-59013-0_17.

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

Enhancing the Cyber Resilience of Sea Drones

Erwin Orye, Maj.
Centre for Digital Forensics
and Cyber Security
Tallinn University of Technology
Tallinn, Estonia
erwin@orye.eu

Gabor Visky
Centre for Digital Forensics
and Cyber Security
Tallinn University of Technology
Tallinn, Estonia
gabor.visky@taltech.ee

Alexander Rohl
School of Computer
and Mathematical Sciences
Faculty of Sciences, Engineering and
Technology
University of Adelaide
Adelaide, Australia
alexander.rohl@adelaide.edu.au

Olaf Maennel
School of Computer
and Mathematical Sciences
Faculty of Sciences, Engineering and
Technology
University of Adelaide
Adelaide, Australia
olaf.maennel@adelaide.edu.au

Abstract: Sea drones are unmanned vessels that operate on or below the water's surface. During the military conflict between the Russian Federation and Ukraine, the latter has demonstrated how to use sea drones to attack Russian targets efficiently. However, as Russia's defences against drone attacks are continuously increasing, the cyber resilience of sea drones is becoming increasingly important. Technological developments in shipping have brought new cybersecurity challenges. This paper contributes to the knowledge on augmenting the cyber robustness of maritime autonomous surface-floating and subaqueous drones. Firstly, we aim to support manufacturers in building affordable sea drones that reduce the cyberattack surface of commercial drones. Secondly, we offer guidance for tactical military commanders on the potential cyber weaknesses in a sea drone's specific operational environments and its reliance on particular technologies. We propose eight distinctive threat categories for cyberattacks against autonomous vessels: attacks to disrupt radio frequency signals; attacks to deceive or degrade sensors; attacks to intercept or modify communications; attacks on operational technology systems; attacks on information technology systems; attacks on artificial intelligence (AI) used for autonomous operations; attacks through supply chains; and attacks through physical access. We use the STRIDE (spoofing,

tampering, repudiation, denial of service, elevation of privilege) [1] methodology in the context of each threat scenario, formulate mitigation measures to reduce the risk for each category, and link methods of cyberattack to each category.

Keywords: *cybersecurity, autonomous, threat modelling, unmanned, vessels, sea drones*

1. INTRODUCTION

Automation, and consequently limited human interaction, has created new vectors for cyberattacks. Cybersecurity is a critical issue for ships with some level of autonomy because of their increased dependence on information and communication technologies (ICT) for ship control, their advanced integration of control systems, their increased connectivity with shore control centres, and their accessibility to (and *from*) the Internet [2].

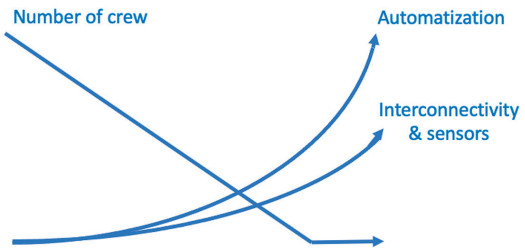
The coexistence of crewed and autonomous vessels (sea *drones*) necessitates the shared use of maritime, canal, and riverine domains. Ensuring the harmonious integration of these two naval transportation modes is vital to the sustainable and effective functioning of waterborne transportation systems.

Industry and academia have conducted extensive research and development in the field of autonomous vessels, such as Wärtsilä's IntelliTug [3], YARA Birkeland [4], L3Harris maritime autonomous systems [5], and Japan's fully autonomous ship program MEGURI2040 [6]. Research projects conducted in academia include, among many others, the University of Plymouth's Cetus Project [7], the Norwegian University of Science and Technology's Autoferry Project [8], and Heli by Tallinn University of Technology and the University of Tartu [9].

Sea drones rely entirely on digital systems with no physical crew to override them. Hence, the consequences of those digital systems being compromised can be more severe than would otherwise be the case.

Figure 1 depicts the evolution of growing automation. In particular, it shows how further automation is possible even when a vessel is already crewless, driven by the need for onshore supervision to become less involved.

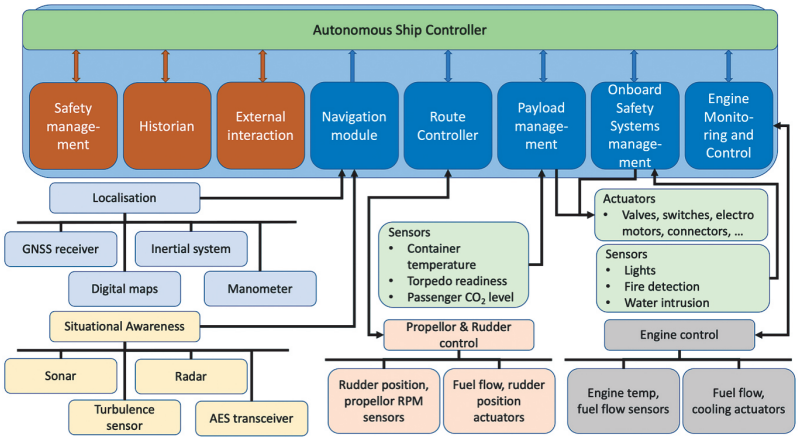
FIGURE 1: NUMBER OF CREW VERSUS THE LEVEL OF AUTONOMY AND RELIANCE ON AN INCREASED NUMBER OF INTERCONNECTED SENSORS



Sea drones come in many different configurations: surface and submarine, commercial and military, large and small, remote-controlled and auto-navigating, and many more [10]. Each configuration is suitable for a specific mission. Vessels can operate for days, weeks, and even longer without human intervention. For example, Saildrone's newest robotic ocean explorer sea drone draws its power from wind and can spend up to 12 months at a stretch out at sea [11]. The US Navy has recently received a prototype ship that can operate autonomously at sea for up to 30 days [12]. And, in 2022, the Nippon Yusen Kabushiki Kaisha (NYK Line) Designing the Future of Full Autonomous Ships (DFFAS) project achieved a 40-hour long autonomous trip across 790 kilometres (491 miles) at sea without human intervention for 99% of the journey [13].

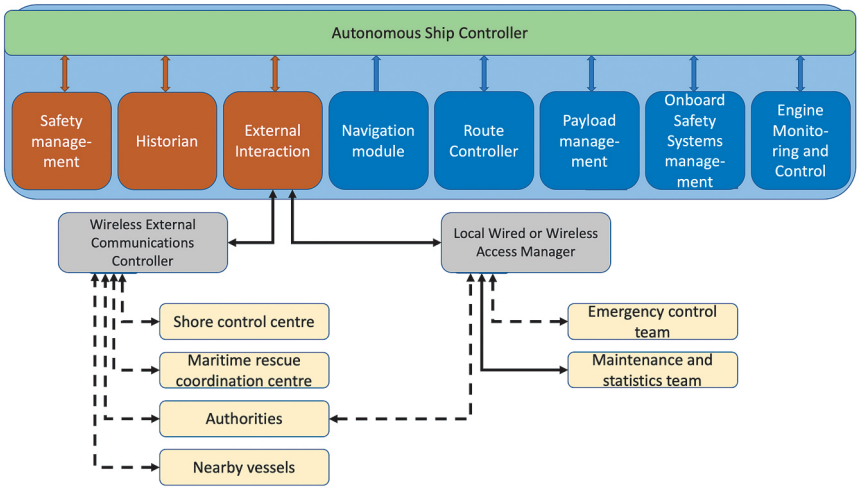
Although the configurations of sea drones might differ, their logical architecture often has the same functionalities. Figure 2 gives an overview of standard sea-drone functionalities.

FIGURE 2: SCHEMATIC OVERVIEW OF THE LOGICAL FUNCTIONS OF AN AUTONOMOUS SEA DRONE



Autonomous sea drones do, at some points, interact with humans, even if rarely or with only a minor impact on their functioning. Figure 3 shows the potential ways in which humans can interact with sea drones. Autonomous vessels have a command-and-control (C&C) channel to execute remote control commands, report sensor statuses, and receive mission instructions from the home base. This C&C channel is not necessarily always active, and the autonomous vessel might have to operate for long periods without supervision, potentially at a considerable physical distance from the control centre. As such, a sea drone needs to be equipped to operate in various uncontrolled environments and for different durations.

FIGURE 3: HUMAN INTERACTION WITH A SEA DRONE AND THE C&C LINKS FOR COMMUNICATIONS



2. RELATED WORK

To our knowledge, a combined study that jointly models the cyber threats, attacks, and defence methods regarding sea drones is not available in the literature (Section 2.B). It is this gap that motivated our research, in which we apply the STRIDE methodology (described in Section 2.A) to real scenarios to identify the adverse effects of cyber threats and the potential methods to defend against them.

A. Literature Review

Silverajan et al. [14] identify seven main attack surfaces through which attackers can gain access to or disrupt operations on uncrewed ships: positioning systems, sensors, firmware, voyage data recorders, intra-vessel networks, vessel-to-land communication,

and remote operations systems. They also define six attack methods: code injection, tampering/modification, positional data spoofing, Automated Identification System (AIS) data spoofing, signal jamming, and link disruption/eavesdropping. However, their work does not cover contextual attack scenarios, possible consequences, or the mitigations required for defence.

Along similar lines, Agamy [15] proposes that the following three threats can affect the cybersecurity of autonomous ships: malicious components added to control systems during building or maintenance sessions, compromised communication links, and position data spoofing. Agamy also discusses a number of examples and regulatory frameworks, such as the International Safety Management Code (ISM), the International Ship and Port Facility Security Code (ISPS), the EU's General Data Protection Regulation (GDPR), and the Australian Cyber Security Center's Final Security Strategy. However, these frameworks do not offer any technical defence measures.

As for the cybersecurity risk assessment of autonomous ships, Tam and Jones [16] model risks relating to the systems and components of autonomous vessels – for example, AIS, Global Navigation Satellite Systems (GNSS), automated mooring systems, cargo management systems, radar, sensors, and voyage data recorders (VDR) – from the perspectives of theft, damage, denial of service, obfuscation, and misdirection. Their model-based framework for maritime cyber-risk assessment (MaCRA) risk model provides a comprehensive method for assessing risk, but the paper does not cover mitigation for the risks or defensive methods against them.

Kavallieratos et al. [17] analyse an autonomous ship into 14 systems: Engine Automation, Bridge Automation, Shore Control Centre, Autonomous Engine Monitoring and Control, Engine Efficiency, Maintenance Interaction, Navigation, Autonomous Ship Controller, Human-Machine Interface, Remote Manoeuvring Support, Emergency Handling, AIS, ECDIS, and Global Maritime Distress and Safety. They then identify threat scenarios for each system using the STRIDE framework. In subsequent research, Kavallieratos and Katskas [18] extend this approach by considering further components of the ship's systems, such as collision avoidance, RADAR, closed circuit television (CCTV), advanced sensor modules, and autopilot systems. These papers give an overview of the risk assessment of autonomous ships. However, they do not detail attack scenarios and defensive measures.

Sungbaek et al. [19] identify cyber threats against autonomous ships, but they do not structure this content into a framework.

B. Threat Model

Threat modelling identifies and enumerates potential security threats and categorizes countermeasures by priority so as to reduce security risks to an acceptable level for the system owner. It includes several safety-focused risk management methodologies for Industrial Control Systems [20]. The CIA-triad (confidentiality, integrity and availability) has been used as a conceptual model in computer security for several decades [21]. The STRIDE methodology, as defined by Shostack [22], categorizes threats corresponding to cybersecurity goals by incorporating three more elements: authentication, non-repudiation, and authorization. The STRIDE threat categories are as follows [23]:

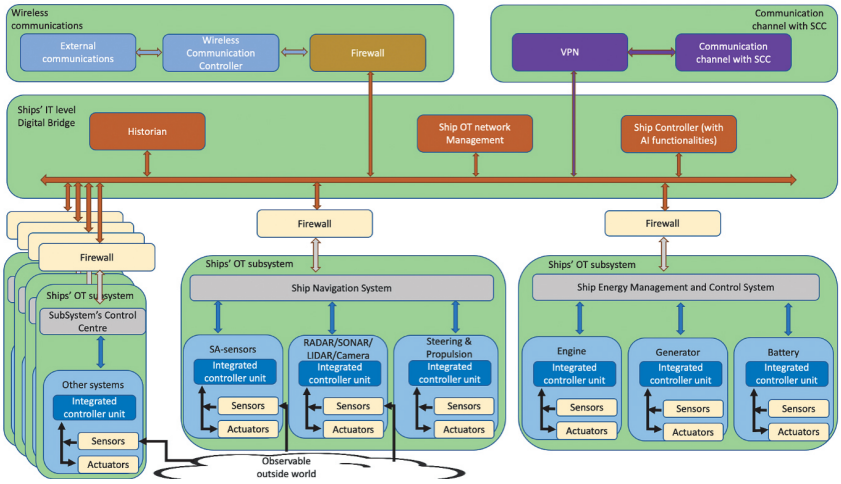
1) Spoofing is the ability of an adversary to masquerade as someone or something else.
2) Tampering refers to modifying or disrupting a system's disk, network, or memory.
3) Repudiation relates to threats where someone denies having taken specific actions that impact the system's operation or disclaims responsibility for the resulting outcomes.
4) Information disclosure involves exposing confidential information to unauthorized individuals.
5) Denial of service refers to compromises to the system's availability that work by consuming the necessary resources for its proper operation.
6) Elevation of privilege refers to situations in which an adversary can execute unauthorized actions.

According to Kim et al., the STRIDE methodology can be used for threat modelling against a distributed control system (DCS) [24]. Since our research focuses on sea drones, and since these are considered a system of DCSs [25], we adopt and use the STRIDE methodology. In that light, our research examines the different possible attacks so as to address the potential threats posed by malicious actors. Instead of focusing on a specific technology used in a particular ship, this paper employs general but transferrable abstractions. Thus, we offer a future-proof approach that can accommodate the broad functionalities of sea drones and cyberattack vectors.

3. RESULTS

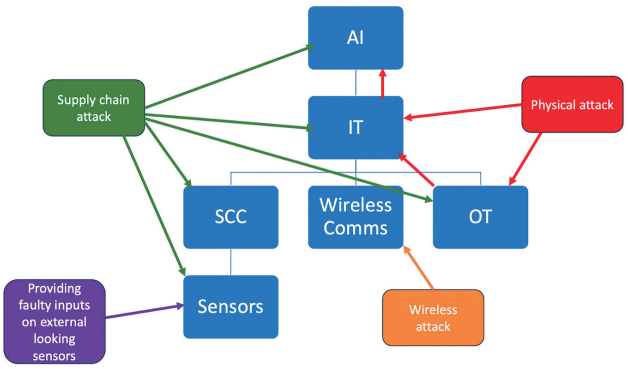
This section introduces the selected attack scenarios and their STRIDE analyses. To help motivate these scenarios, we must first consider the necessary functions of sea drones and their related subsystems. Figure 4 shows an abstract schematic overview, focusing only on the different types of equipment and how they relate to each other within an autonomous vessel.

FIGURE 4: SCHEMATIC OVERVIEW OF SUBSYSTEMS IN AN AUTONOMOUS VESSEL



To understand the attack surface of these subsystems, we must examine the lines for information flow. Figure 5 gives an overview of each sea-drone subsystem's possible attack vectors.

FIGURE 5: THE POSSIBLE ATTACKS ON A SEA DRONE'S SUBSYSTEMS



Taking these attacks and the interaction of the different subsystems as our starting point, we defined eight areas that we examine in more detail through the STRIDE methodology:

1) Attacks to intercept, modify or disrupt wireless communications
2) Attacks to deceive or degrade sensors
3) Attacks on operational technology (OT) systems
4) Attacks on information technology (IT) systems
5) Attacks on artificial intelligence (AI) for autonomous operations
6) Supply chain attacks (SCA)
7) Physical attacks to launch cybersecurity attacks and
8) Attacks against the shore control centre (SCC)

A. Attacks to Intercept, Modify, or Disrupt Wireless Communications

Description: RF signals serve various purposes in relation to wireless communication, radar systems, and other wireless technologies. Disrupting RF signals involves actions taken to interfere with or disturb these signals. This can be accomplished through various means – for instance, jamming, interference, or deliberate manipulation of the signals – that degrade or turn off communication between devices or systems that rely on these signals.

Possible scenario: A design weakness, implementation, or design flaw in authentication or encryption can lead to signal manipulation.

S: Communication protocols such as AIS are easy to spoof in the maritime sector [26].
T: An attacker on the wireless C&C channel between the control station and the autonomous vessel could take over complete control of the ship.
R: There is often at lack of robust resilience against data modification within existing RF protocols. The absence of features to facilitate repudiation becomes apparent.
I: Autonomous vessels have sensors onboard. Some vessels provide some information on the fly through wireless channels to the home base. Access by third parties to sensitive information can lead to the disclosure of information.
D: Disruption of the C&C channel can lead to the vessel being made idle or execute fully automated actions, such as return to base. In any case, it is likely to lead to a denial of service for the operation of the vessel.
E: Accessing the C&C channel can allow deeper access to the system and overruling immutable parameters from a distance.

Possible mitigations:

- Using inertial systems or recognizing the environment with sensors and correlation with databases can mitigate incorrect GNSS input data or the unavailability of GNSS input data.
- There are multiple mitigations to protect against jamming, such as channel hopping, spectrum spreading, MIMO (multiple-input and multiple-output) based mitigation, channel coding, rate adaptation, and power control [27].
- A VPN solution or similar can potentially protect the C&C channels themselves and add additional authentication and integrity checks such as counters on messages, structure of messages, digital signatures, and so on.
- Communications that rely on interoperability – for instance, a communication channel between harbour and vessel, AIS, weather forecast broadcast, GNSS, or GDMSS – are vulnerable to attacks by design. However, there are possible countermeasures. For example, the autonomous vessel could try to filter out fake AIS messages by looking at the physical layer of the message and correlating this with previous messages to compensate. On the other hand, ignoring AIS messages too readily might decrease the vessel's situational awareness, which can increase the danger of collisions. Securing those channels would be the next level of security for autonomous ships.

B. Attacks to Deceive or Degrade Sensors

Description: The sensors that capture information outside an autonomous ship offer a high-privilege way for attackers to influence the ship's operation because the attacker does not need physical access to the vessel to compromise these. In this regard, the location or proximity of the vessel is a condition to consider.

Another attack would be fooling internal sensors such as fire detection, engine failure, stability sensors, and so on. However, this would require first gaining physical access and initiating attacks on the sensors from there.

Possible scenario: A ship's sensors are prone to jamming and the injection of false echoes. The same applies to sensors designed for very short-distance situational awareness, such as cameras and illuminating LEDs on optical sensors.

S: An attack that changes the vessel's surroundings so that the sensors pick up a modified input. If an attacker knows a sensor's behaviour, they can modify the input so as not to trigger attention from the digital bridge.
T: Sensors need calibration before use. An attacker tampering with calibration (e.g., for a depth sensor) might cause severe havoc.

R: Most attacks that fool the sensors and provide erroneous information are challenging to repudiate.
I: Knowing how many sensors and what characteristics they have might indicate what type of vessel it is and how to attack it.
D: Ensuring that sensors cannot provide measurements in their everyday working range would constitute a denial of service for those sensors.
E: Attacks against sensors do not necessarily provide a means for privilege escalation.

Possible mitigations:

- The autonomous ship should have sufficient sensors based on entirely different technologies, compare the inputs from those sensors, and make decisions based on as complete information as possible. The greater the range of different technologies installed, the more difficult it becomes for the attacker to successfully provide all of the wrong inputs simultaneously. For example, using lidar, radar, and AIS systems to determine if the vessel is on a collision course with another ship is more reliable than using only AIS or only one radar sensor. In the former situation, hackers might need to intervene in close proximity to the targeted ship to influence its behaviour. Good situational awareness of the vessel's surroundings, above and under the sea level, is vital to detecting any signs of an intruder. The correlation of inputs from different sensors and specific sensors over time can reveal threats.
- Log files and histograms might help the digital bridge determine if any sensors are producing incorrect input data and take action to mitigate the problem. Such action can equally help with faulty sensors when there is no intervention from a malicious actor.

C. Attacks on Operational Technology Systems

Description: Most of the digital components of an autonomous ship are operational technology (OT) systems. Traditionally, protocols used in OT systems are vulnerable to various cyberattacks since there is no standard encryption mechanism implemented in most communication protocols, and the authentication happens at the hardware level or not at all. For example, all major fieldbus protocols – such as Modbus, DNP3, Profinet and EtherCAT – lack authentication or encryption. Thus, if they manage to get access to the network, attackers can disrupt network operations or manipulate I/O messages to cause a failure in the control process [28].

Possible scenario: Different attacks are possible in this context, such as first hacking the C&C link and, with privilege escalation, getting into the core networks. Gaining

physical access to the system, such as through maintenance ports or even the physical wires, is another option.

S: It is straightforward to spoof an endpoint in an OT network since there is no authentication, and, therefore, it is easy to spoof an existing hardware address.
T: OT systems are prone to supply chain attacks and insider threats. For example, maintenance personnel could constitute an insider threat. An example of the former would be if a manufacturer or another actor in the supply chain of the OT endpoint or core element were to reveal undocumented functionalities that an attacker could use to launch an attack.
R: There are often no logfiles for OT networks since the total number of messages is substantial, even though each individual message might be small in size.
I: An attacker can read all the information passing on the bus. Depending on the size and type of endpoints, they can map the topology of the network and the functionalities of each endpoint.
D: By flooding the bus with messages, the denial of service of an endpoint becomes straightforward. If the endpoint is only sending information, this information is not reaching any destination. If the endpoint reads information from the bus, it will not receive any helpful input data.
E: The OT systems are often at the heart of the autonomous vessel. Protection focuses on threats from the outside. An attacker might try to go from the OT network (or bus) to get to a central controller and from there to the digital bridge.

Possible mitigations:

- By segregating networks, the amount of helpful information available on any one segment can be limited. Gateways, firewalls, and other security measures are essential to reduce the risk of an attacker gaining access to more segments, controllers, or even the digital bridge.
- Considerations should be made for implementing enhanced security for control systems, encrypting all volatile and non-volatile memory, securing bus protocols between different devices, and segregating/segmenting the networks with controls. Implementing these measures is challenging because of the number of OT devices on board and the need for common relevant standards.
- Another possible line of defence is to analyse all traffic in real time with anomaly-detector machine-learning algorithms that can identify abnormal behaviour.

D. Attacks on Information Technology Systems

Description: Attacks on information technology (IT) systems modify the firmware of various components and devices on autonomous ships, operating systems, and software running on higher-level machines.

Possible scenario: With an attack on the IT systems, an attacker gains access to the digital bridge. Depending on the elevation of privilege on the IT system(s), this might allow them to gain complete control over the vessel.

S: Without firmware integrity verification and authorization for firmware updates, an attacker could perform an unauthorized firmware update. If the operator activates this option, the attacker could execute this via maintenance interfaces and over-the-air updates.
T: Malicious firmware updates can tamper with the functionalities of the autonomous vessel.
R: Without signed versions of software updates, it is nearly impossible to attribute an attack digitally.
I: Once an attacker is in the IT systems, they might have access to databases, (sensor) data, localization, the health status of the vessel, and other critical information.
D: When the central IT system is not responding as designed, the autonomous vessel is no longer executing its mission.
E: One of the most effective paths for an attacker of an IT system is an escalation of privilege. To gain complete control over the autonomous vessel, the attacker needs access to many functionalities in the IT system.

Possible mitigations:

- The first question that an operator of an autonomous vessel should decide upon is whether software or firmware updates are allowed over the air. Depending on the situation, one option will be better than the other. If operators at the SCC do not have access to the ship when they discover a significant software flaw, one option is to implement a patch immediately over the air. Still, enabling this access increases the attack surface for attackers. It is essential to know the status of the software and hardware and, therefore, use signed versions of firmware from trusted companies, define policies on who, when, and how to update the system, and, last but not least, test the software for functionality and security before installing it.
- Preferential redundancy is critical for making autonomous decisions. Use equipment and software from different vendors that provide the same functionality to install multiple independent calculation chains and, ultimately, use a voting system that decides what action to take.

E. Attacks on AI for Autonomous Operations

Description: Machine learning code provides functions that can replace the human factor. This kind of software is, therefore, interesting from an attacker's point of view since it is directly engaged with the decision-making process. Attacks on machine learning software aim to cause misjudgement or malfunction.

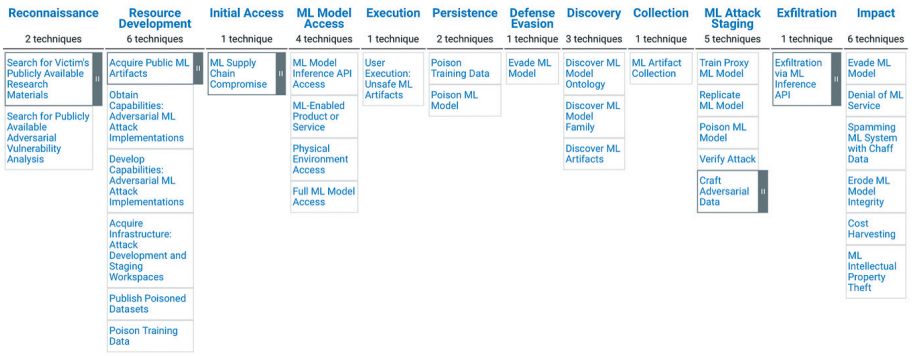
Possible scenario: Typical attacks on machine learning include evasion attacks (to fool a machine learning model by corrupting the query), model poisoning, and data pointing.

S: An attack on specific sensors might change the input for the AI coming from that sensor and fool the algorithms into changing the outcomes of decisions.
T: Modifying the behaviour of the AI software can result in different responses to sensor inputs. If the attacker has enough knowledge about the vessel, they might use this to execute actions on the ship.
R: Without digital signatures to allow changes in the AI software, other traces are required to achieve repudiation, which can be challenging.
I: Tampering with the AI system might lead to the full disclosure of all data available or generated on the vessel.
D: When altering the AI system, it is possible to achieve a complete denial of service of the autonomous vessel by spoofing input values to the AI that take unusually long to process.
E: Given that this attack targets the data of the AI system, it is important to note that it does not facilitate privilege escalation.

Possible mitigations:

- Select training datasets that focus on how to work effectively under sensor degradation or actuator failures. It is also crucial to consider what happens if the opponent knows the algorithms or the learning datasets and can create special conditions by fooling some sensors. Machine learning could help discover weaknesses in other machine learning software.
- Figure 6 shows all the attacks against deployed machine-learning systems according to the ATLAS framework.
- Extensive testing in extreme conditions should be conducted. The datasets for learning the system should include ways of responding to cyberattacks.

FIGURE 6: MITRE'S ADVERSARIAL THREAT LANDSCAPE FOR ARTIFICIAL INTELLIGENCE SYSTEMS [29]



F. Supply Chain Attacks

Description: Attacks on the supply chain – which can have various sources, including third-party vendors, internal employees, and others – include disruption of operations, compromise of sensitive information, financial losses, reputational damage, legal and regulatory implications, and so on [30].

Possible scenario: A hacker steals a certificate used to vouch for the legitimacy or safety of a company's product, or a hacker leverages the tools for building software applications to introduce security weaknesses in the development process. Similarly, preinstalled malware can represent a valid threat scenario – for instance, if there is a malicious component in the firmware.

S: Implement faulty MAC addresses, ID numbers, or other mechanisms to receive information from the internal bus or networks.
T: Malicious code or components can be injected into the product through targeted attacks that initialize, for example, communication to a C&C server, thus creating a tampering backdoor into the system.
R: It is always difficult to tell which actor implemented a backdoor, a spy module, a modified firmware, and so on. Was it the chip manufacturer, the print board, the integrator, the shipping company, or other stakeholders?
I: When malicious actors trick individuals, a phishing attack can lead to information disclosure or compromised security, sometimes providing access with elevated rights.
D: A compromised component can cause a denial of service on an autonomous ship.
E: Malicious code running in a software component with elevated privileges can offer access to the IT systems with elevated rights.

Possible mitigations:

- Considering the multiple forms they can take, defending against SCA requires a range of different techniques, including auditing the IT (shadow) infrastructure, a highly secure build and update infrastructure, up-to-date software assets, application of client-side protection solutions, and so on [31].
- It is necessary to precisely follow up on all modifications made to a product, from designing to manufacturing integration to decommissioning.

G. Physical Attacks to Launch Cybersecurity Attacks

Description: If an autonomous ship operates in the open sea, physical protection for the vessels can easily be weaker than otherwise.

Possible scenario: Various maintenance interfaces on autonomous ships, such as USB, Serial, JTAG and RJ45, could be exploited as initial attack vectors. Even if there are physical locks to prevent unauthorized physical access to these interfaces, there is a possibility that an attacker could compromise the locks and make unauthorized connections through these interfaces as the autonomous ship navigates in the open sea for an extended period of time.

S: With physical access to the vessel, an attacker gains an entry point to the digital systems without facing the difficulties of accessing interface points with the outside world. It makes sense that those interface points are the way in with the least privilege and the most extensive logging. Otherwise, determining the ease of spoofing the system depends on the exact location of the entry point.
T: There are many possible tampering actions, from swapping disks to plugging USB sticks with malware into maintenance. Different attacks are possible depending on the time available, size, computational power, design, and complexity.
R: Physical attacks are complicated to attribute digitally. Forensics might find some artefacts if, for example, malware leaves some digital traces.
I: Information disclosure is a risk for all internal communication that is not encrypted and where the attacker with physical access can extract the data. The same goes for databases that contain unencrypted data.
D: All physical destruction – for instance, unplugging a cable or flooding a data bus – will lead to denial of service of parts or the whole of the autonomous vessel.
E: An attacker still has to achieve elevation of privilege unless they can physically replace the IT system with their own.

Possible mitigations:

- There are many options to reduce the risk of physical access and the impact of such an attack: segregation and segmentation of the networks, cable fault sensors that detect anomalies, sensors that raise the alarm on intrusion, external sensors such as drones or satellites that surveil the neighbourhood of the vessel, physical protection measures such as locks to reduce the chances of obtaining physical access, firewalls between segments, time scheduled maintenance slots, and so on.
- Cost, the attacker's benefit, the vessel's value, available space, allowed weight, power consumption, and so on will probably determine the number and type of countermeasures that can and need to be put in place.

H. Attacks Against the Shore Control Centre to Launch a Cyberattack on a Sea Drone

Attack description: Most autonomous ships have a C&C channel to receive input from the home base. This communication can be sporadic when tasking a mission to remote control with some automatic functions. The shore control centre (SCC) has a privileged entry point to the vessel from the outside. Access to the SCC might compromise one or more ships.

Possible scenario: Inappropriate segregation between the C&C network and the office network at the SCC or inappropriate control over removable media/mobile devices might compromise the C&C network, which can result in the transmission of unauthorized commands to autonomous ships or disruption of the C&C communication channel itself.

S: When instructions come from a hacker that spoofs the SCC – if the attacker has the encryption key for the VPN tunnel to the vessel, for example – the vessel will be unable to differentiate between legitimate and spoofed instructions.
T: The attacker can install malware through the C&C channel or modify the vessel's behaviour if remote updates are allowed.
R: If the attacker leaves traces in the SCC, it is possible to attribute an attack, but the traces of the login on the vessel will not help identify an attacker if the messages are well crafted.
I: The hacker will have access to all the data the SCC has access to. For example, if a vessel sends observations from its sensors directly to the SCC, the attacker will receive the same information.
D: An attack against the SCC does not necessarily lead to a denial of service for an autonomous vehicle. However, because of the level of automation, it still poses a danger.

E: Once the attacker can take over the C&C control channel, they might still need an elevation of privilege for the functionalities the SCC cannot execute from a distance. The SCC retains a large number of permissions to intervene when unexpected situations occur.

Possible mitigations:

- The SCC is a typical IT infrastructure with specific software to create instructions for the vessel and communicate this in a particular way. Therefore, the protection of the SCC is most similar to protection measures implemented by banks or for critical infrastructure. ISO27K series, National Institute of Standards and Technology (NIST), or similar guides the management of cybersecurity risks in this field.

4. DISCUSSION

Cybersecurity relates to risk assessment. Criminals attacking cargo vessels do not have the same profile as state actors who also show interest in specialized military, research, and governmental-operated vessels. Configurations of such specialized vessels can differ extensively in terms of the type and number of sensors, redundancy of subsystems, processing power, machine learning algorithms, and many other features. Thus, not all the subsystems previously mentioned need to be present, and the size and number of existing subsystems can differ significantly.

What actions a system owner takes to reduce the impact of cybersecurity attacks depends on the threat scenario, the residual risk an operator wants to assume, the threat level, the importance of the mission, their finances, and the time they have available to operationalize a vessel. Improving cybersecurity boils down to securing the complete software and hardware supply chain. Early levels of indicators of compromise (IOCs) and intelligence about advanced persistent threats (APTs) are a significant help when it comes to being informed about the threat scenario and level.

Complete autonomous ship operations have a larger cybersecurity attack surface. Still, depending on the setting, this can be acceptable since such ships have the advantage that there will be no loss of life and no way to demand ransoms when something happens to the vessel and non-existent crew.

Verification at different levels is essential to reducing the risk of the vessel being compromised:

1) Identification and authentication control: Who is allowed to access the system, and can you verify that this person is who they claim to be?
2) User control: Who is allowed to execute which commands?
3) Integrity control: Are you sure that the instructions have not been tampered with?
4) Data confidentiality control: Are you sure that adversaries cannot intercept information?
5) Restricted: Ensure everyone has access to information only on a need-to-know basis. This concept is very crucial with regard to insider threat issues.

Following our STRIDE analysis of the eight subsystems, a sea drone owner or manufacturer should take the relevant steps to improve the cyber resilience of their sea drone:

1) Analyse the system into its logical components according to Figure 4.
2) Define all the data fluxes between each system component and the external world.
3) Identify threats for each system component and function based on the operational use of the sea drone and the corresponding attackers' profiles.
4) Once the threats for each system component are identified, the STRIDE model indicates where vulnerabilities might arise. Software exists to support the technical process of finding specific vulnerabilities. For example, the Microsoft Threat Modelling Tool (MTMT) [32] implements the STRIDE framework at the software level. Open-source software, such as the open software templates building tool, inserts STRIDE threats in the generated template by searching common vulnerabilities and exposures (CVE) databases [33].
5) Take mitigation measures such as controlling information flows, adapting policies and installing control mechanisms. Implement effective mitigation strategies based on the specific discovered vulnerabilities.

5. CONCLUSIONS AND FUTURE WORK

Our research identified potential threats against autonomous maritime vehicles and provided a framework for their mitigation. Following that, we used the STRIDE attack model to highlight the cybersecurity aspects of sea drones and considerations

relevant to those, thus providing a solid background for manufacturers and end users willing to improve their sea drones.

We provided a framework and inventory of cyber risks for the engineers who develop sea drones and the users of sea drones. While we did not focus on the different components or parts of the sea drones, we grouped these into general but applicable subsystems to provide a foundational path towards developing detailed solutions for a specific sea drone. In our judgement, this approach fits the field best since each sea drone is a system of systems with its own individual specialized configuration.

Our research was limited to autonomous sea drones and crewed ships, depending on the level of automation. Although we focused only on technology-related measures, training people and improving processes are similarly crucial to cyber defence.

Many sea drones will soon serve as military [34] and merchant ships [35]. Our research aims to help industry and policymakers create a global ecosystem for safe and secure autonomous shipping.

REFERENCES

[1] L. Kohnfelder and P. Garg, 'The threats to our products', Microsoft Security Development Blog, 1999. [Online]. Available: https://www.microsof t.com/security/blog/2009/08/27/the-threats-to-our-products/0Ahttps://adam.shostack.org/microsoft/The-Threats-To-Our-Products.docx

[2] S. K. Katsikas, 'Cyber security of the autonomous ship', in *CPSS 2017 – Proceedings of the 3rd ACM Workshop on Cyber-Physical System Security, Co-located with ASIA CCS 2017*, pp. 55–56, 2017.

[3] 'Initial sea trials successfully completed by Wärtsilä & PSA Marine's ground-breaking "IntelliTug" project'. Wärtsilä. Accessed: Jan. 3, 2024. [Online]. Available: https://www.wartsila.com/media/news/13-03-2020-initial-sea-trials-successfully-completed-by-wartsila-psa-marine-s-ground-breaking-intellitug-project-3290931

[4] 'Yara Birkeland'. Yara. Accessed: Jan. 3, 2024. [Online]. Available: https://www.yara.com/news-and-media/media-library/press-kits/yara-birkeland-press-kit/

[5] 'Autonomous systems'. L3Harris. Accessed: Jan. 3, 2024. [Online]. Available: https://www.l3harris.com/all-capabilities/autonomous-systems

[6] 'The Nippon Foundation Meguri2040 fully autonomous ship program'. Nippon Foundation. Accessed: Jan. 3, 2024. [Online]. Available: https://www.nippon-foundation.or.jp/en/what/projects/meguri2040

[7] 'Uncrewed surface vessel (USV) Cetus'. University of Plymouth. Accessed: Jan. 3, 2024. [Online]. Available: https://www.plymouth.ac.uk/research/esif-funded-projects/usv-cetus

[8] 'Autoferry'. NTNU. Accessed: Jan. 3, 2024. [Online]. Available: https://www.ntnu.edu/autoferry

[9] 'Scientists launch Estonia's first autonomous maritime research vessel'. ERR. Accessed: Jan. 3, 2024. [Online]. Available: https://news.err.ee/1609117841/scientists-launch-estonia-s-first-autonomous-maritime-research-vessel

[10] N. Klein, D. Guilfoyle, M. S. Karim, and R. McLaughlin, 'Maritime autonomous vehicles: New frontiers in the law of the sea', *International and Comparative Law Quarterly*, vol. 69, no. 3, pp. 719–734, 2020.

[11] 'Saildrone launches a 72-foot autonomous seabed-mapping boat'. TechCrunch. Accessed: Jan. 11, 2024. [Online]. Available: https://techcrunch.com/2021/01/11/saildrone-launches-a-72-foot-autonomous-seabed-mapping-boat/?guccounter=2

[12] 'The navy's new autonomous ship can run by itself for 30 days'. Accessed: Jan. 11, 2024. [Online]. Available: https://www.popularmechanics.com/military/navy-ships/a43033206/navy-ship-can-operate-autonomously-for-30-days/

[13] ‘Autonomous cargo ship completes 500 mile voyage, avoiding hundreds of collisions’. Electrek. Accessed: Jan. 10, 2024. [Online]. Available: https://electrek.co/2022/05/13/autonomous-cargo-ship-completes-500-mile-voyage-avoiding-hundreds-of-collisions/
[14] B. Silverajan, M. Ocak, and B. Nagel, ‘Cybersecurity attacks and defences for un-manned smart ships’, in *Proceedings – IEEE 2018 International Congress on Cybermatics: 2018 IEEE Conferences on Internet of Things, Green Computing and Communications, Cyber, Physical and Social Computing, Smart Data, Blockchain, Computer and Information Technology, iThings/Gree*, pp. 15–20, 2018.
[15] K. S. M. Agamy, ‘The impact of cybersecurity on the future of autonomous ships’, *International Journal of Recent Research in Interdisciplinary Sciences*, vol. 6, no. 2, pp. 10–15, 2019.
[16] K. Tam and K. Jones, ‘Cyber-risk assessment for autonomous ships’, in *2018 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*.
[17] G. Kavallieratos, S. Katsikas, and V. Gkioulos, ‘Cyber-attacks against the autonomous ship’, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11387, pp. 20–36, 2019.
[18] G. Kavallieratos and S. Katsikas, ‘Managing cyber security risks of the cyber-enabled ship’, *Journal of Marine Science and Engineering*, vol. 8, no. 10, pp. 1–19, 2020.
[19] S. Cho, E. Orye, G. Visky, and V. Prates, *Cybersecurity Considerations in Autonomous Ships*. Tallinn: CCDCOE, 2022.
[20] H. Abdo, M. Kaouk, J.-M. Flaus, and F. Masse, ‘A safety/security risk analysis approach of industrial control systems: A cyber bowtie—combining new version of attack tree with bowtie analysis’, *Computers & Security*, vol. 72, pp. 175–195, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404817301931
[21] M. Whitman and H. Mattord, *Principles of Information Security*. Boston, MA: Cengage Learning, 2021. [Online]. Available: https://books.google.ee/books?id =Hwk1EAAAQBAJ
[22] A. Shostack, *Threat Modeling*. Nashville, TN: John Wiley & Sons, 2014.
[23] J. Meier, A. Mackman, S. Vasireddy, M. Dunner, R. Escamilla, and A. Murukan, *Improving Web Application Security*. Microsoft Corporation, 2003. [Online]. Available: https://www.microsoft.com/en-us/download/confirmation.aspx?id=1330
[24] K. H. Kim, K. Kim, and H. K. Kim, ‘STRIDE-based threat modeling and DREAD evaluation for the distributed control system in the oil refinery’, *ETRI Journal*, vol. 44, no. 6, pp. 991–1003, Nov. 2022, doi: 10.4218/etrij.2021-0181.
[25] K. Tam, K. Forshaw, and K. Jones, ‘Cyber-SHIP: Developing next generation maritime cyber research capabilities’, in *Conference Proceedings of ICMET Oman*, Muscat, Oman, Nov. 2019, doi: 10.24868/icmet.oman.2019.005.
[26] ‘Spoofed warship locations—automatic identification system (AIS)’. Popular Mechanics. Accessed: Jan. 7, 2024. [Online]. Available: https://www.popularmechanics.com/military/navy-ships/a37261561/ais-ship-location-data-spoofed/
[27] H. Pirayesh and H. Zeng, ‘Jamming attacks and anti-jamming strategies in wireless networks: A comprehensive survey’, *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 767–809, 2022.
[28] E. D. Knapp and J. T. Langill, *Industrial Network Security: Securing Critical Infrastructure Networks for Smart Grid, SCADA, and Other Industrial Control Systems*, 2nd ed. Waltham, MA: Syngress, 2015.
[29] ‘ATLAS’. MITRE. 2021. [Online]. Available: https://atlas.mitre.org
[30] H. S. Berry, ‘The importance of cybersecurity in supply chain’, in *2023 11th International Symposium on Digital Forensics and Security (ISDFS)*, IEEE, May 2023, doi: 10.1109/ISDFS58141.2023.10131834.
[31] ‘What are supply chain attacks? Examples and countermeasures’. Fortinet. Accessed: Jan. 7, 2024. [Online]. Available: https: //www.fortinet.com/resources/cyberglossary/supply-chain-attacks
[32] ‘Microsoft threat modeling tool’. Microsoft. Accessed: Jan. 3, 2024. [Online]. Available: https://learn.microsoft.com/en-us/azure/security/develop/threat-modeling-tool
[33] M. Da Silva, M. Puys, P. H. Thevenon, S. Mocanu, and N. Nkawa, *Automated ICS template for STRIDE Microsoft Threat Modeling Tool* (ACM International Conference Proceeding Series), 2023.
[34] ‘US Navy aims to field manned-unmanned fleet within 10 years’. Defense News. Accessed: Jan. 3, 2024. [Online]. Available: https://www.defensenews.com/naval/2023/04/12/us-navy-aims-to-field-manned-unmanned-fleet-within-10-years/
[35] Z. H. Munim and H. Haralambides, ‘Advances in maritime autonomous surface ships (MASS) in merchant shipping’, *Maritime Economics and Logistics*, vol. 24, no. 2, pp. 181–188, 2022, doi: 10.1057/s41278-022-00232-y.

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

Defeating and Improving Network Flow Classifiers Through Adversarial Machine Learning

Yannick Merkli
LatticeFlow
Zurich, Switzerland
ymerkli@latticeflow.ai

Roland Meier
armasuisse Science and Technology
Cyber-Defence Campus
Thun, Switzerland
roland.meier@ar.admin.ch

Martin Strohmeier
armasuisse Science and Technology
Cyber-Defence Campus
Thun, Switzerland
martin.strohmeier@ar.admin.ch

Vincent Lenders
armasuisse Science and Technology
Cyber-Defence Campus
Thun, Switzerland
vincent.lenders@ar.admin.ch

Abstract: Recent work has shown that machine learning models can be vulnerable to an adversary crafting targeted inputs designed to cause mispredictions. This is critical in security-related applications such as network intrusion detection systems. While past attacks such as mimicry or gradient-based attacks are able to efficiently generate adversarial examples, they require potentially large input modifications, which is not effective at defeating network flow classifiers.

In this work, we show that small modifications to the input (e.g., the traffic that the attacker generates) are enough to manipulate the outcome of a classifier. We focus on minimally evasive adversarial examples to defeat tree-ensemble-based network flow classifiers. We develop an attack that builds on a previous attack introduced by Kantchelian et al. in 2016, which formulates evasion for tree ensembles as a Mixed Integer Linear Program, and which we extend by supporting discrete and categorical features, implementing per-feature evasion costs and modeling inter-feature dependencies. This makes our attack more applicable to the network flow classification problem, which typically uses diverse and interdependent input features.

We demonstrate our attack on the network flow classifier developed by Känzig et al. in 2019, which was trained to detect command and control (C&C) channels in the

Locked Shields cyber defense exercise. Our evaluation shows that minor perturbations of 1 to 4 flow features suffice to successfully fool the classifier. We further retrain the network flow classifier using state-of-the-art adversarial boosting and robust decision tree training published by Chen et al. in 2019. For example, using adversarial boosting, the resulting robust classifier shows a 62.5% increased median evasion distance while achieving equivalent precision and recall on unperturbed samples as the original classifier.

Keywords: *adversarial machine learning, traffic classification, Locked Shields*

1. INTRODUCTION

Machine learning (ML) algorithms are being applied to an ever-growing number of security-sensitive domains, such as malware detection or network intrusion detection [1]–[3]. These algorithms have proven capable of finding novel patterns and handling amounts of data that are not processable by humans, which is very beneficial in data-intensive applications. For instance, network intrusion detection systems (NIDS) protect networks by monitoring traffic and detecting anomalous behavior. Traditionally, these systems relied on experts developing a set of rules that encode known malicious behavior. However, with networks growing larger and more heterogeneous and network traffic showing higher variability and much larger volume, this approach is becoming increasingly challenging. Furthermore, this approach is unable to discover previously unknown attacks and is limited by the fact that most network traffic today is encrypted. For these reasons, researchers have proposed ML-based NIDS and have shown these to be capable of quickly and accurately identifying malicious behavior [1], [4], [5].

However, security-sensitive domains have an intrinsically adversarial nature, with adversaries actively trying to bypass any protective measures that have been put in place. Recent work on adversarial machine learning suggests that many learning algorithms are vulnerable to input perturbations. Such perturbations can happen randomly (e.g., because the environment changed) and lead to significantly degraded classification performance, as shown by Gehri et al. [6]. However, such perturbations can also happen because an adversary deliberately introduces them to cause mispredictions [7]–[10].

Successful attacks on security-sensitive ML systems can have disastrous consequences. For instance, botnets are one of the most serious threats in today's

security environment, causing malicious activities such as distributed denial of service (DDoS) attacks. Successfully defending against botnets requires efficient and accurate detection of botnet traffic, which is made challenging because of the obfuscation and resilience techniques employed by attackers. ML-based botnet detection systems are thus becoming increasingly popular and have proven to perform well in correctly labeling botnet communication [1], [2].

In this work, we focus on defeating and improving a tree-ensemble-based network flow classifier developed by Känzig et al. [1] that was trained for command and control (C&C) flow classification. In order to evaluate the classifier's robustness in an adversarial setting, we adopt the role of an attacker trying to evade the flow classifier to examine how a C&C flow needs to be modified if the classifier is not to label the resulting flow as malicious. Past approaches have focused on mimicking a known benign sample [11], [12], which requires potentially large adversarial modifications, leading to suboptimal malicious behavior. Instead of taking this approach, we focus on optimal evasion, i.e., finding the minimal adversarial modification needed to evade the classifier. This approach leads to optimal behavior for the attacker and allows for making exact guarantees of the classifier's robustness. We leverage an existing adversarial attack [7] and extend this to the network intrusion detection domain. We then apply our attack to Känzig et al.'s [1] network flow classifier and show that minor perturbations of 1 to 4 flow features suffice to successfully evade the classifier. Finally, we improve the flow classifier by applying state-of-the-art adversarial defenses.

The paper is organized as follows: Section 2 describes the background to the work described in this paper. Section 3 summarizes related work. Section 4 presents the design of our attack framework. Section 5 details the results of the evaluation. In Section 6, we conclude and discuss future research directions.

2. BACKGROUND

In this section, we briefly introduce the background to the work described in this paper.

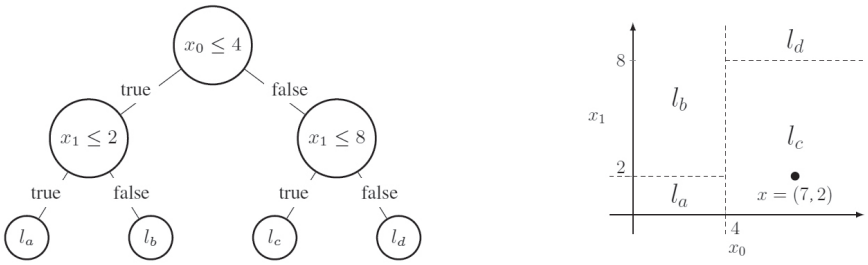
A. Decision Tree Ensemble

A decision tree consists of nodes, where each leaf node holds a prediction score, and each non-leaf node holds a logical decision rule on a given input feature and has two outgoing edges that are labeled *true* and *false*, indicating the decision path depending on the rule evaluation for an input sample x. When given an input sample, the prediction value is found by walking the tree from root to leaf such that an edge is in the decision path if and only if the associated decision rule evaluates to *true* for the

input sample. For instance, consider the simple decision tree in Figure 1 and a sample $x = (7,2)$. The decision rule $x_0 \leq 4$ evaluates to false and $x_1 \leq 8$ to true, thus the active leaf for sample x is leaf l_c. A decision tree ensemble consists of a set of decision trees, where its prediction is an aggregation of the predictions of each active leaf in each individual tree in the ensemble.

Since decision trees naturally encode human-interpretable decision rules, a key advantage of tree-based learning algorithms is interpretability. Tree ensemble classifiers are widely used in practice, especially in the security domain, where they have been shown to be superior compared to other learning algorithms, such as support vector machines (SVM) and neural networks [1], [2].

FIGURE 1: AN EXAMPLE OF A SIMPLE DECISION TREE AND ITS FEATURE SPACE PARTITION. LEAF l_c WOULD BE ACTIVE FOR SAMPLE $x = (7,2)$



B. Network Traffic Classification

Network traffic consists of bidirectional flows, whereby the term flow refers to a set of packets that are sent from a source to a destination and that share a common set of properties. Most commonly, a flow is defined by its 5-tuple (source and destination IP addresses and transport-layer ports and the transport-layer protocol).

In order to classify network flows, one has to extract a representative set of features. The most commonly used features are statistical flow features (e.g., the average packet size), which can be extracted from packet traces.

C. Mixed Integer Linear Programming

A mixed integer linear programming (MILP) problem deals with a mathematical optimization problem that consists of solely linear functions and a finite set of discrete or continuous variables. A MILP problem consists of an objective, a set of constraints and a set of variables, where the optimal solution optimizes the objective while fulfilling the set of constraints.

Since most problems cannot be translated directly into an objective function and a set of constraints, solving a problem using MILP requires a careful modeling and translation process. Once the problem has been modeled into a MILP formulation, a global optimum can be found through the use of an efficient solver.

D. Locked Shields

Locked Shields is a complex international cyber defense exercise organized annually by the NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) in Tallinn, Estonia [13]. Locked Shields focuses on realistic scenarios and simulates the full complexity of a large-scale cyber incident, including regular business IT, critical infrastructure, military systems, strategic decision-making, and legal and communication aspects. The exact scenario varies from year to year, but the exercise is generally organized as a real-time Red team vs. Blue team exercise in which the Blue teams are the training audience.

E. The Target Classifier

The classifier we analyze in this paper was developed by Känzig et al. [1] and is designed to quickly and reliably identify C&C channels in the setting illustrated in Figure 2. It uses a random forest classifier with 128 trees of maximum tree depth 10, was trained on data from the Locked Shields exercises from 2017 and 2018, and uses the 10 or 20 most important features (depending on the classifier version) listed in Table I. In the following, *top10* and *top20* will refer to classifiers that were trained on the 10 or 20 most important features, respectively.

FIGURE 2: NETWORK APPLICATION DOMAIN WITH THE DEPLOYED FLOW CLASSIFIER AS LIVE DEFENSE AGAINST A BOTMASTER THAT COMMUNICATES WITH INFECTED ENDHOSTS VIA A C&C CHANNEL

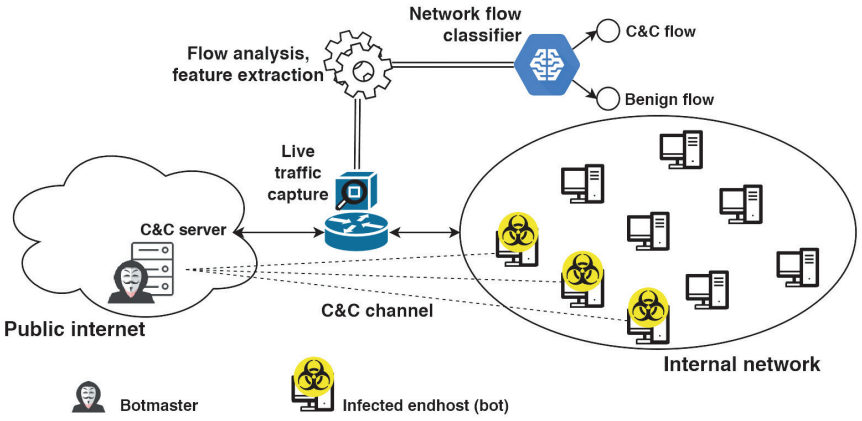


TABLE I: FEATURES ON WHICH THE TARGET FLOW CLASSIFIERS ARE TRAINED. THE TOP10 AND TOP20 CLASSIFIERS ARE TRAINED ON FEATURES 1–10 AND 1–20, RESPECTIVELY. *Bwd* INDICATES TRAFFIC FROM THE C&C SERVER DESTINED FOR AN INFECTED ENDHOST

ID	Feature name	Description
1	Protocol	The transport layer protocol
2	dstIntExt	Internal or external dstIP
3	Active Mean	Mean time a flow was active before becoming idle
4	Init Fwd Win Byts	Total number of bytes sent in the initial window in the forward direction
5	FIN Flag Cnt	Number of packets with FIN
6	Bwd Pkt Len Min	Minimum size of packet in the backward direction
7	Flow Pkt/s	Number of flow packets per second
8	Fwd IAT Max	Maximum time between two packets sent in the forward direction
9	Flow IAT Mean	Mean time between two packets sent in the flow
10	Subflow Fwd Pkts	The average number of packets in a subflow in the forward direction
11	Flow IAT Max	Maximum time between two packets sent in the flow
12	Fwd IAT Tot	Total time between two packets sent in the forward direction
13	Subflow Bwd Pkts	The average number of packets in a subflow in the backward direction
14	Subflow Fwd Byts	The average number of bytes in a sub flow in the forward direction
15	Bwd Header Len	Total bytes used for headers in the backward direction
16	Tot Bwd Pkts	Total packets in the backward direction
17	Fwd Pkt Len Std	Standard deviation size of packet in the forward direction
18	Fwd Seg Size Min	Minimum segment size observed in the forward direction
19	Bwd Pkt Len Std	Standard deviation size of packet in the backward direction
20	Bwd IAT Mean	Mean time between two packets sent in the backward direction

3. RELATED WORK

There is a large body of work applying machine learning algorithms to network intrusion detection. Existing approaches use algorithms such as support vector machines (SVM) [14], k-nearest neighbors [15], neural networks [16]–[18], or decision trees [1], [5], [19].

Several recent works have investigated the robustness of machine learning algorithms towards adversarial examples, which are targeted perturbations of an original sample that change the prediction of a model [7], [9], [10], [20]. Such adversarial attacks have shown that, generally, small targeted perturbations suffice to fool a model. Meanwhile, various adversarial defenses have been proposed that attempt to defend against adversarial examples by incorporating them during training [7], [8], [21].

In our work, we extend an existing adversarial attack [7] to the network intrusion detection setting and apply it to a tree-ensemble-based network flow classifier [1]. Further to that, we use adversarial training [7], [8] to improve the adversarial robustness of the flow classifier.

4. ATTACK FRAMEWORK

We now provide an overview of our attack framework and explain the relevant technical details. While our attack framework is designed for the network intrusion detection problem, the framework is generic and can also be applied to machine learning systems in other domains.

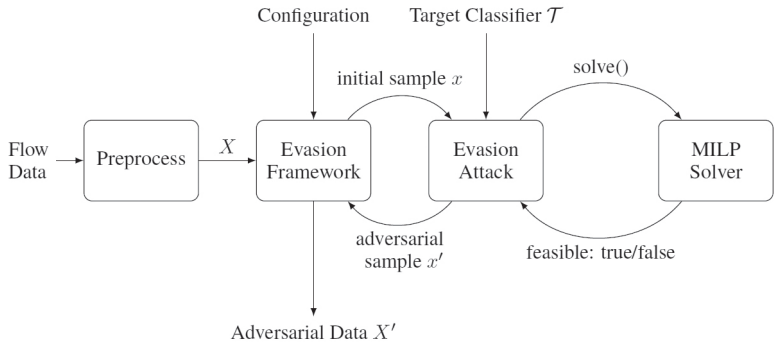
A. Overview

Our attack implements an automatic, end-to-end evasion framework, meaning that given an initial set of samples and a target tree-ensemble classifier, the system runs an optimal evasion attack for every sample and returns the adversarial samples. We designed our attack framework for the C&C channel evasion scenario. This scenario consists of a botmaster that tries to spread commands to infected endhosts inside a network that is protected by the target classifier. The botmaster's goal is to minimally adjust its network flows so that the classifier does not label them as malicious. In order to give strong security guarantees, we assume that the attacker has white-box access to the classifier. However, we assume that the attacker cannot modify the classifier's training.

Our attack framework is illustrated in Figure 3. The system first performs any necessary preprocessing on the raw flow data, which gives the input data X. The

evasion framework then receives both the input data X and a configuration that defines the feature space of X as well as feature constraints and costs. The evasion framework then runs the attack on X, receiving an adversarial sample for each input sample. The attack needs to first initialize by parsing the target classifier and translating its structure into a MILP problem. By constructing an evasion framework around the attack that passes a set of input samples X, we avoid having to repeatedly perform this translation, which makes the attack more efficient and, thus, more realistic for use in real network environments.

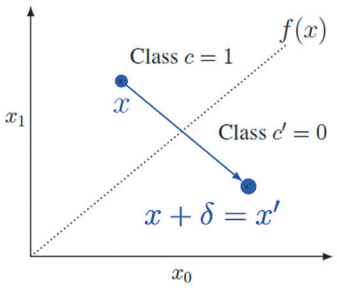
FIGURE 3: ATTACK FRAMEWORK



B. Attack

In abstract, the problem of evading a classifier can be described as follows: Given a classification algorithm $f: \mathcal{X} \to \mathcal{Y}$ and an input $x \in \mathcal{X}$ with $y = f(x)$, we want to find an adversarial input $x' \in \mathcal{X}$ that is assigned class $y' = f(x') \neq y$. Finding any input x' that is classified differently than the initial sample x is generally trivial and not particularly useful to the attacker. Therefore, what we want to find is an adversarial input $x' = x + \delta$ such that δ is minimized (cf. illustration in Figure 4).

FIGURE 4: EVASION OF AN INITIAL INPUT $x \in \mathcal{X}$ FOR A CLASSIFICATION ALGORITHM $f: \mathcal{X} \to \mathcal{Y}$



The attack we use in our work is based on the tree ensemble attack by Kantchelian et al. [7], which minimizes a loss function $\mathcal{L}: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$:

$$\min_{x' \in \mathcal{X}} \mathcal{L}(x, x') \text{ such that } f(x) \neq f(x')$$

While Kantchelian et al. propose to use any l^p-norm $\mathcal{L}^p(x, x') = \left(\sum_{i=1}^{d} |x_i - x_i'|^p\right)^{1/p}, \forall p \in \{1, 2, \ldots, \infty\}$ as their loss function, we generalize this loss function with a linear adversarial cost function [22]. This allows us to model the cost of adapting a network flow feature on a per-feature base by specifying a linear factor $a_j \in \mathbb{R}_+$ for each feature $j \in [0, d-1]$. Moreover, we are further able to capture immutable features by setting $a_j \rightarrow \infty$. Our new proposed loss function $\mathcal{L}$ for the adversarial attack is then given as:

$$\mathcal{L}^p(x, x') = \left(\sum_{j=0}^{d-1} a_j \left|x_j - x_j'\right|^p\right)^{1/p}, \forall p \in \mathbb{N}_+$$

and

$$\mathcal{L}^\infty(x, x') = max_j\left(a_j \left|x_j - x_j'\right|\right), j \in [0, d-1]$$

Feature Constraints: Finding an adversarial input x' according to the problem statement given above would require that every feature $x_j, j \in [0, d-1]$ could be modified independently; however, it is overly optimistic to assume that this can be done. In domains such as network intrusion detection, features typically have dependencies on each other. For instance, consider the *FIN Flag Cnt* and *Protocol* features of the target classifier. The *FIN Flag Cnt* feature can be 0 or 1 for TCP but has to be 0 for UDP since UDP does not have FIN flags. Thus, an adversarial sample that modifies *FIN Flag Cnt* from 0 to 1 for a UDP flow does not correctly model the given feature dependencies, and the botmaster will not be able to send a network flow that corresponds to the computed adversarial sample. Furthermore, feature constraints give the attacker more precise control over the adversarial flow since any custom constraints can be specified.

In order to model potentially complex feature constraints, we propose two types of constraints that can be imposed on the input features.

Simple constraints model any feature constraint of the form:

$$x_j' \circ t_j, j \in [0, d-1], t_j \in \mathbb{R}$$

Where $\circ$ is a binary operator $\circ \in \{<, >, \leq, \geq\}$, and t_j is any threshold on feature x_j'. This constraint forces feature x_j' of the evasion sample to take a value for which $x_j' \circ t_j$

evaluates to true. This allows us to model simple specifications on the evasion sample, such as constraining an evasion sample to have UDP as its protocol.

Dependent constraints model any implication feature constraint of the form:

$$x'_i \circ_i t_i \Rightarrow x'_j \circ_j t_j, i, j \in [0, d-1], t_i, t_j \in \mathbb{R}$$

Where $\circ_i$,$\circ_j$ are binary operators $\circ_i$,$\circ_j \in \{<, >, \leq, \geq\}$, and t_i, t_j are any thresholds on features x'_i, x'_j, respectively. This constraint enforces that if $x'_i \circ_i t_i$ evaluates to true for feature x'_i of the evasion sample x', then $x'_j \circ_j t_j$ must evaluate to true for feature x'_j of the evasion sample x'. This allows us to model inter-feature dependencies, such as enforcing that if the evasion sample has Protocol UDP, then *FIN Flag Cnt* has to be 0.

Furthermore, each constraint can either be global or per sample. Global constraints are applied once when initializing the attack framework. Per-sample constraints are applied and removed for every input sample, which allows unique constraints to be specified for every sample $x \in \mathcal{X}$.

C. Adversarial Training

In Section 5, we show that our attack on C&C classifiers is able to efficiently find adversarial samples, which can be used by a botmaster to send commands to infected endhosts without getting detected by the C&C classifier. Following these findings, we will show that the classifier's robustness can be improved by using state-of-the-art adversarial training methods. First, we use the adversarial boosting method proposed by Kantchelian et al. [7], which efficiently generates adversarial samples that can be incorporated when training the classifier. Further to that, we use robust decision tree training [8] in order to build a robust boosting tree classifier based on the model parameters of the initial classifier.

5. RESULTS

We evaluate our attack on four different versions of the target classifier, where the classifiers were trained on data from Locked Shields 2017 and 2018 (from here on referred to as *LS17* and *LS18*) and on top10 and top20, respectively. We instantiate the attack with the $\mathcal{L}^1$ loss and analyze for each adversarial input $x' = x + \delta$ the required adversarial perturbation $|\delta|_1$ and the susceptibility of each feature—i.e., how often a certain feature is adapted in order to evade the classifier. Further to that, we use the following fixed dependent feature constraints to model the feature dependency between the *Protocol* feature and the *FIN Flag Cnt and the Init Fwd Win Byts* features.

- $x_{Protocol} \geq 12 \Rightarrow x_{FINFlagCnt} < 0.5$: UDP does not have FIN flags
- $x_{Protocol} \geq 12 \Rightarrow x_{InitFwdWinByts} < 0$: UDP does not have an initial TCP window size. Flowmeter sets the Init Fwd Win Byts feature to -1 for UDP flows

Note that, due to these constraints, the optimal evasion attack rarely modifies the *Protocol* feature, even though we set the evasion cost of the *Protocol* feature to 1. Due to the feature dependencies of the *Protocol* feature, changing the protocol requires modifying the *Init Fwd Win Byts* feature, which results in large evasion distances.

Finally, all optimal evasion attacks were executed on a server with 10 Intel Xeon E5-2699 v3 cores at 2.30 GHz and 16 GB of RAM, and we used Gurobi [23] as a MILP solver.

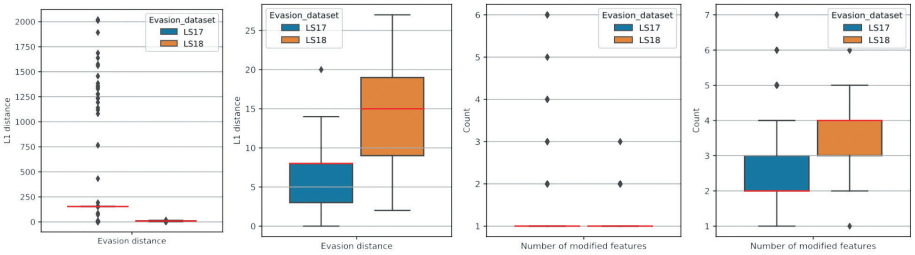
A. Flow Classifier Evasion

We first show the results for running our attack on all four considered classifier versions. For each evaluation, we run our attack on 1,000 samples and show the ℓ_1 evasion distances and the number of features that had to be modified. Figure 5b.1 shows that for the top10 classifiers, modifying one feature suffices to successfully evade the classifier in almost all cases. Table II shows that for the LS17 classifier, the most commonly modified feature is the *Init Fwd Win Byts* feature and that it must be decreased by 154 bytes on average. This is a minor modification, considering that the initial TCP window byte size lies in the region of [0.65535]. For the LS18 classifier, the most frequently modified feature is the *Subflow Fwd Pkts* feature, and we see that if this feature is modified, the number of subflow packets in the forward direction must be decreased by 9 packets on average. One could argue that sending fewer packets per flow is a more restrictive modification since the botmaster might need a minimum amount of information transmitted. However, this can be addressed by opening multiple flows and sending fewer packets per flow.

For the top20 classifiers, when we look at Table III and Figure 5b.2, we see similarly that a small subset of features is particularly susceptible. Compared to the top10 classifiers, the median number of modified features is larger for the top20 classifiers, namely 2 (LS17) and 4 (LS18)—this is as expected since the top20 classifiers use double the number of features. The most frequently modified features are *Tot Bwd Pkts, Subflow Bwd Pkts, Bwd Header Len*, and *Subflow Fwd Pkts*, which can all be modified by the adversary. Table III shows how much a given feature has to be modified in order to evade the classifier. For example, considering the two most commonly used evasion features, *Tot Bwd Pkts and Subflow Bwd Pkts*, we see that it suffices to send just 2–4 fewer packets to successfully evade the classifier.

We can thus conclude that an adversary can evade all flow classifiers by modifying a small subset of flow features. We further note that the most commonly modified features (*Init Fwd Win Byts, Subflow Fwd Pkts, Subflow Bwd Pkts, Tot Bwd Pkts, and Bwd Header Len*) can all be controlled by the adversary exactly.

FIGURE 5: EVASION DISTANCE AND THE NUMBER OF MODIFIED FEATURES WHEN RUNNING OUR ADVERSARIAL ATTACK ON 1,000 SAMPLES. IN A.1 AND B.1, THE RESULTS FOR LS17 ARE ON THE LEFT-HAND SIDE, AND THE RESULTS FOR LS18 ARE ON THE RIGHT-HAND SIDE



(a.1) top10 classifiers (a.2) top20 classifiers (b.1) top10 classifiers (b.2) top20 classifiers

TABLE II: PER-FEATURE EVASION DISTANCES OF THE TOP10 FLOW CLASSIFIERS. THE EVASION DISTANCE MEDIAN AND STANDARD DEVIATION ARE ONLY TAKEN ON NON-ZERO EVASION DISTANCES DUE TO THE SPARSITY OF THE DATA. FEATURES THAT WERE MODIFIED LESS THAN TEN TIMES COMBINED ARE NOT SHOWN

Feature	LS17			LS18		
	evasion dist. median	evasion dist. std	# of times modified	evasion dist. median	evasion dist. std	# of times modified
Subflow Fwd Pkts	-1.0	258.839	55	-9.0	2.982	958
Init Fwd Win Byts	-154.0	20.729	925	0	0	0
Fwd IAT Max	1.0	128.783	40	1.0	0.0	30
Flow Pkts/s	-57.683	854.448	27	-2.329	2.379	11
Flow IAT Mean	-414.159	239.196	18	-2.572	2.725	14
Bwd Pkt Len Min	10.0	3.374	23	7.0	2.070	8
Active Mean	8.0	6.397	16	0	0	0

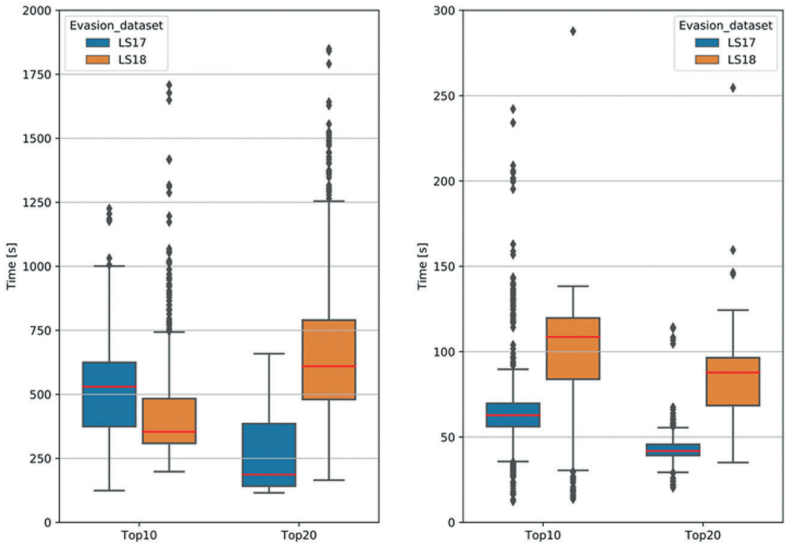
TABLE III: PER-FEATURE EVASION DISTANCES OF THE TOP20 FLOW CLASSIFIERS. THE EVASION DISTANCE MEDIAN AND STANDARD DEVIATION ARE TAKEN ONLY ON NON-ZERO EVASION DISTANCES DUE TO THE SPARSITY OF THE DATA. FEATURES THAT WERE MODIFIED LESS THAN 100 TIMES COMBINED ARE NOT SHOWN

Feature	LS17			LS18		
	evasion dist. median	evasion dist. std	# of times modified	evasion dist. median	evasion dist. std	# of times modified
Tot Bwd Pkts	-3.0	1.640	879	-4.0	0.991	951
Subflow Bwd Pkts	-3.0	1.640	877	-2.0	0.966	952
Bwd Header Len	1.0	1.299	118	-6.0	4.375	864
Subflow Fwd Pkts	1.0	0.877	210	-8.0	4.050	651
Fwd Seg Size Min	7.0	1.815	108	7.0	1.705	161
Bwd Pkt Len Std	0.049	0.807	108	-8.605	-	1

1) Evasion Time

MILP solvers can be computationally expensive, which is a deciding factor with respect to the practical relevance of our attack. For instance, considering the Locked Shields exercise, an optimal evasion attack that takes several hours or days is not practical since the exercise only takes place over a time span of four days. To demonstrate the practical relevance of our attack, we analyze the evasion time when using Gurobi [23] and PuLP [24]. Figure 6 shows the time to run the attack for a single sample. The median evasion time is between 3 and 10 minutes using PuLP and between 1 and 2 minutes using Gurobi. Thus, while Gurobi solves the optimal evasion problem significantly faster, both cases are clearly practical since evasion times in the range of several minutes do not pose a major limitation for an attacker.

FIGURE 6: PER-SAMPLE EVASION TIME USING PuLP AND GUROBI AS MILP SOLVER, 500 SAMPLES



(a) PuLP (b) Gurobi

2) Adversarial Transferability

Past work [20], [25] suggests that adversarial samples are transferable between different models, meaning that an adversarial input created for one classifier has a high likelihood of being adversarial for a different classifier too.

We evaluate adversarial transferability by checking whether adversarial samples for LS17 classifiers are also adversarial on the respective LS18 classifiers and vice-versa. As Table IV shows, adversarial samples from our attack are less transferable compared

to other attacks and learning algorithms, such as gradient descent attacks and neural networks. This result is expected since our attack is highly targeted, and the minimal evasive sample is set right at the decision boundary of the classifier. A slight variation in decision boundary can thus push an adversarial sample back into a malicious decision region, which in turn gives a true positive classification.

TABLE IV: TRANSFERABILITY OF ADVERSARIAL SAMPLES GENERATED BY THE OPTIMAL EVASION ATTACK

Classifier	**Recall** (%) on LS18 adversarial samples	**Classifier**	**Recall** (%) on LS17 adversarial samples
LS17 Top10	96.2	LS18 Top10	84.0
LS17 Top20	79.9	LS18 Top20	51.7

3) Comparing Optimal and Approximate Evasion

In addition to the optimal evasion attack, Kantchelian et al. also propose a faster approximate attack. In the following, we compare the adversarial inputs found by the approximate and the optimal attack. We run 1,000 evasions for every classifier version and limit the number of coordinated descent iterations to 30 in order to break exceedingly long coordinate descent searches. The results of these evaluations are shown in Table V. If we compare the evasion distances of adversarial inputs found by the approximate attack and the optimal attack, we can clearly see that the approximate attack finds adversarial inputs that require significantly larger modifications.

B. Adversarial Training

In this section, we use the adversarial training approaches discussed previously to train potentially more robust network flow classifiers, and we evaluate the resulting classifiers in terms of their classification performance and robustness against our attack.

1) Adversarial Boosting

We use adversarial boosting, as proposed by Kantchelian et al., to improve the robustness of the target classifiers. If we look at Table V, we see that the approximate attack finds adversarial samples with significantly larger perturbations and has a low success rate. Therefore, we decided not to use approximate attacks for adversarial boosting but instead to use our optimal attack to generate large sets of adversarial samples.

In our evaluation, we ran 10 concurrent evasion attacks on the top20 LS17 classifier over the course of 17 days, which is the computation time required by the MILP solver. This allowed us to generate 20,0152 adversarial samples, which corresponds to 16.1% of the number of malicious samples in the Locked Shields 2017 dataset. We

then trained the model proposed by Känzig et al. on a combined dataset consisting of the original Locked Shields 2017 samples and the generated adversarial samples. We first evaluated the classification performance of the resulting robust classifier on the Locked Shields 2018 dataset. Table VI shows the classification performance of the retrained classifier on the LS18 dataset compared to the original classifier. Compared to the original top20 LS17 classifier by Känzig et al., the robust classifier's classification performance is largely equivalent and even performs slightly better in malicious class recall.

We further evaluated the robustness of the retrained classifier using our optimal evasion attack. If we look at Figure 7a.1, we see that the median evasion distance for the adversarial retrained top20 LS17 classifier is 13.0. This corresponds to a 1.625× increase in robustness compared to the original top20 LS17 classifier, which has a median evasion distance of 8.0 (see Figure 5a.1). If we look Figure 7a.2, we further see that the median number of modified features for the retrained top20 LS17 classifier is equal to 2, which is equivalent to the original top20 LS17 classifier.

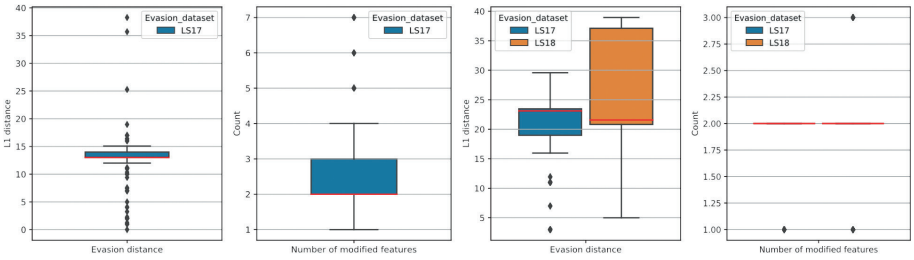
2) Robust Boosting Tree

We trained a gradient-boosting tree model using the robust tree training framework [8] for both the Locked Shields 2017 and 2018 datasets and for the top20 feature sets. Once again, we measured the robustness of each trained classifier by performing an optimal evasion attack for 1,000 samples and compared this to the robustness of the initial classifier.

TABLE V: APPROXIMATE EVASION EVALUATIONS FOR ALL CLASSIFIER VERSIONS. *EVASION DISTANCE INCREASE* REFERS TO THE INCREASE COMPARED TO THE OPTIMAL EVASION DISTANCE OF THE RESPECTIVE CLASSIFIERS

Classifier	ℓ_1 evasion distance median	evasion distance increase	# modified features median	evasion time median	evasion success rate (%)
LS17 Top10	35449.649	230.2×	2.0	3.164s	84.0
LS18 Top10	4260.424	473.4×	2.0	44.628s	15.2
LS17 Top20	2048.250	256.0×	5.0	13.429s	69.1
LS18 Top20	50410.264	3360.7×	6.0	15.390s	71.2

FIGURE 7: EVASION DISTANCE AND THE NUMBER OF MODIFIED FEATURES FOR THE ADVERSARIALLY TRAINED CLASSIFIERS. IN FIGURE 7B.2, THE RESULTS FOR LS17 ARE ON THE LEFT-HAND SIDE AND THE RESULTS FOR LS18 ARE ON THE RIGHT-HAND SIDE



(a.1) (a.2) (b.1) (b.2)

(a) Adversarial Boosting (b) Robust Boosting Tree

Table VII shows the classification performance of the robust top20 classifiers compared to the original top20 classifiers. We observe that the classification performance for the *benign* class decreases only slightly for both robust classifiers, except for the *benign* precision of the robust top20 LS18 classifier, which increases by 1.5%. For the robust LS17 classifier, *malicious* recall decreases by 42% and all averaged classification metrics decrease by roughly 4% compared to the original top20 LS17 classifier. The robust LS18 classifier performs 5.7% worse in *malicious* precision. However, it has an 11.3% increase in malicious recall, and thus an increase of around 1% for all averaged classification metrics, compared to the original top20 LS18 classifier.

We further compare the robustness of the robust classifiers to the original classifiers. Figure 7b.1 shows the evasion distance of the robust top20 classifiers. The robust top20 LS17 classifier has a median evasion distance of 23.1, which corresponds to a 2.89× increase in robustness compared to the original top20 LS17 classifier, which has a median evasion distance of 8.0. The robust top20 LS18 classifier has a median evasion distance of 21.6, which corresponds to a 1.44× increase in robustness compared to the original top20 LS18 classifier, which has a median evasion distance of 15.0. If we look at Figure 7b.2, we see that the number of modified features is 2 for the robust top20 LS17 classifier, which is equivalent to the original top20 LS17 classifier. The robust top20 LS18 classifier has only 2 modified features per evasion, compared to 4 for the original top20 LS18 classifier. Thus, while the evasion distance of the robust top20 LS18 classifier does increase, the decreased number of modified features per evasion can also be seen as a decrease in robustness.

TABLE VI: CLASSIFICATION PERFORMANCE OF THE ROBUST RETRAINED AND THE ORIGINAL KÄNZIG ET AL. TOP20 LS17 CLASSIFIER ON THE LS18 DATASET. *BENIGN, MALICIOUS* REFERS TO NON-C&C AND C&C FLOWS, RESPECTIVELY

Class	Precision (%)		Recall (%)		F1-score (%)	
	rob.	orig.	rob.	orig.	rob.	orig.
Benign	99.7	99.7	99.9	99.9	99.8	99.8
Malicious	98.7	98.7	97.6	97.3	98.1	98.0
Overall	99.6	99.6	99.6	99.6	99.6	99.6

TABLE VII: CLASSIFICATION PERFORMANCE OF THE ROBUST RETRAINED AND THE ORIGINAL KÄNZIG ET AL. TOP20 FLOW CLASSIFIERS. *BENIGN, MALICIOUS* REFERS TO NON-C&C AND C&C FLOWS, RESPECTIVELY

Class	**LS17**						**LS18**					
	Precision (%)		Recall (%)		F1-score (%)		Precision (%)		Recall (%)		F1-score (%)	
	rob.	orig.	rob.	orig.	rob.	orig.	rob.	orig.	rob.	orig.	rob.	orig.
Benign	95.3	99.7	99.7	99.9	97.4	99.8	99.4	97.9	98.9	99.8	99.2	98.8
Malicious	95.6	98.7	55.3	97.3	70.1	98.0	92.5	98.2	95.4	84.1	93.9	90.6
Overall	95.3	99.6	95.3	99.6	94.7	99.6	98.6	97.9	98.5	97.9	98.5	97.8

6. CONCLUSION

In this work, we evaluated the robustness of network intrusion detection classifiers by applying an attack to tree-ensemble-based classifiers. We presented an extended attack that builds on a previous attack by Kantchelian et al., and we designed an attack framework that handles end-to-end evasion of input samples. We showed that minor modifications to C&C flows suffice to successfully fool the target classifier and observed that most adversarial modifications can be implemented in practice. Furthermore, we showed that by using state-of-the-art robust training methods, we can increase the robustness of the target classifier.

However, there are limitations and extensions to consider for the future. First, while our work shows that minor modifications to the input features suffice to evade the target classifier, in practice, the attacker needs to be able to map the adversarial input to a network flow. Thus, the attacker needs to send traffic in such a way that the target system's traffic capture and preprocessing actually map the sent network flow to the intended adversarial input. While this is straightforward for features such as the TCP window size or the number of sent packets, it becomes more difficult for timing-based features such as the interarrival time. Thus, a natural extension to our work would be to implement a traffic modification system that leverages our attack framework to find adversarial inputs and then automatically modifies network traffic

to evade the target classifier in real time. Considering our results, the most commonly modified features can directly be modified on an endhost; thus, such a system could be directly implemented as a local system on the endhost or as a proxy. Another possibility would be to implement the system in-network using programmable data planes [26]. Second, we use an l^p-norm with linear costs as the loss function for our attack. However, this loss function does not necessarily reflect an attacker's effective effort to modify a network flow. Consider, for example, the *Init Fwd Win Byts* feature. In our loss function, modifying the *Init Fwd Win Byts* by 20 bytes incurs twice the loss of modifying it by 10 bytes. However, in practice, the attacker's effort for the two modifications is likely very similar. Thus, future work could explore different adversarial loss functions that better model an attacker's effective effort for modifying network traffic.

REFERENCES

[1] N. Känzig, R. Meier, L. Gambazzi, V. Lenders, and L. Vanbever, "Machine learning-based detection of C&C channels with a focus on the Locked Shields cyber defense exercise," in *2019 11th International Conference on Cyber Conflict (CyCon)*, vol. 900, IEEE, 2019, pp. 1–19.

[2] B. Abraham, A. Mandya, R. Bapat, F. Alali, D. E. Brown, and M. Veeraraghavan, "A comparison of machine learning approaches to detect botnet traffic," in *2018 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2018, pp. 1–8.

[3] C. Smutz and A. Stavrou, "Malicious PDF detection using metadata and structural features," in *Proceedings of the 28th Annual Computer Security Applications Conference*, 2012, pp. 239–248.

[4] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," *ACM SIGCOMM Computer Communication Review*, vol. 34, no. 4, pp. 219–230, 2004.

[5] J. Zhang, M. Zulkernine, and A. Haque, "Random-forests-based network intrusion detection systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C* (*Applications and Reviews*), vol. 38, no. 5, pp. 649–659, 2008.

[6] L. Gehri, R. Meier, D. Hulliger, and V. Lenders, "Towards generalizing machine learning models to detect command and control attack traffic," in *2023 15th International Conference on Cyber Conflict: Meeting Reality (CyCon)*, IEEE, 2023, pp. 253–271.

[7] A. Kantchelian, J. D. Tygar, and A. Joseph, "Evasion and hardening of tree ensemble classifiers," in *International Conference on Machine Learning*, 2016, pp. 2387–2396.

[8] H. Chen, H. Zhang, D. Boning, and C.-J. Hsieh, "Robust decision trees against adversarial examples," 2019, *arXiv:1902.10660*.

[9] B. Biggio et al., "Evasion attacks against machine learning at test time," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2013, pp. 387–402.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[11] P. Laskov et al., "Practical evasion of a learning-based classifier: A case study," in *2014 IEEE Symposium on Security and Privacy*, IEEE, 2014, pp. 197–211.

[12] D. Wagner and P. Soto, "Mimicry attacks on host-based intrusion detection systems," in *Proceedings of the 9th ACM Conference on Computer and Communications Security*, 2002, pp. 255–264.

[13] "Locked shields." CCDCOE. 2020. [Online]. Available: https://ccdcoe.org/exercises/locked-shields/

[14] M. S. Pervez and D. M. Farid, "Feature selection and intrusion classification in NSL-LKDD cup 99 dataset employing SVMs," in *8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014)*, IEEE, 2014, pp. 1–6.

[15] H. Shapoorifard and P. Shamsinejad, "Intrusion detection using a novel hybrid method incorporating an improved KNN," *International Journal of Computer Applications*, vol. 173, no. 1, pp. 5–9, 2017.

[16] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A deep learning approach to network intrusion detection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 1, pp. 41–50, 2018.

[17] R. Vinayakumar, K. Soman, and P. Poornachandran, “Applying convolutional neural network for network intrusion detection,” in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, 2017, pp. 1222–1228.
[18] G. Zhao, C. Zhang, and L. Zheng, “Intrusion detection using deep belief network and probabilistic neural network,” in *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, vol. 1, IEEE, 2017, pp. 639–642.
[19] B. Ingre, A. Yadav, and A. K. Soni, “Decision tree based intrusion detection system for NSL-KDD dataset,” in *Information and Communication Technology for Intelligent Systems (ICTIS 2017)*, vol. 2, no. 2, Springer, 2018, pp. 207–218.
[20] C. Szegedy et al., “Intriguing properties of neural networks,” 2013, *arXiv:1312.6199*.
[21] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” 2017, *arXiv:1706.06083*.
[22] D. Lowd and C. Meek, “Adversarial learning,” in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 2005, pp. 641–647.
[23] “Gurobi optimizer reference manual,” Gurobi Optimization LLC, 2020. [Online]. Available: http://www.gurobi.com
[24] S. Mitchell, M. OSullivan, and I. Dunning, “PuLP: A linear programming toolkit for Python,” University of Auckland, Auckland, New Zealand, 2011.
[25] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1765–1773.
[26] P. Bosshart et al., “P4: Programming protocol-independent packet processors,” *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 3, pp. 87–95, 2014.

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)

Artificial Intelligence System Risk Management Methodology Based on Generalized Blueprints

Dan Bogdanov*
Information Security Research Institute
Cybernetica, Estonia

Paula Etti*
Information Security Research Institute
Cybernetica, Estonia

Liina Kamm*
Information Security Research Institute
Cybernetica, Estonia
liina.kamm@cyber.ee

Fedor Stomakhin*
Information Security Research Institute
Cybernetica, Estonia

Abstract: The rapid uptake of artificial intelligence (AI) systems requires similar advances in their governance. Public and private sector institutions want to adopt new AI tools as they perceive potential efficiency gains and value from them. As with every technological advance, the uptake phase of AI is the ideal time to improve the governance, cybersecurity and safety of these systems.

The cybersecurity risks in AI systems are similar to the ones in other information technology systems. However, the regulation of AI systems is changing, so new governance tools are needed. Furthermore, the safety and societal impact of AI depends on the technological choices made when building the systems (e.g., biased training data, overfitted machine learning models, model poisoning attacks or needlessly computation-heavy algorithms).

AI tools built with large language model technology seem to speak our languages and therefore appear deceptively easy to adopt. The goal of our research is to provide risk management tools that are similarly easy to use, even if they later lead the adopter into setting up a full technical quality management system.

We have created three blueprints of AI system deployments to which an organization deploying AI can match their use case. For each blueprint, we have created high-

* Co-funded by the European Union. The views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Cybersecurity Competence Centre. Neither the European Union nor the European Cybersecurity Competence Centre can be held responsible for them. This work was supported by the Estonian Centre of Excellence in AI (EXAI), funded by the Estonian Ministry of Education and Research.

level guidance on which cybersecurity, data rights and ethical aspects the deploying organization needs to consider. Those building AI systems can quickly match their use cases against the blueprints and speed up the secure and ethical adoption of AI.

Keywords: *artificial intelligence, cybersecurity, data governance, data protection, risk management*

1. INTRODUCTION

Artificial intelligence (AI) is developing at a rapid pace and organizations are deploying AI systems to make their processes more efficient. However, the deployment of AI systems brings with it the need for risk management to ensure the deployed systems are secure, safe and compliant with all relevant laws. This can be a difficult task for smaller organizations that do not have a mature risk management framework in place. Even large organizations with an established risk management process must account for AI-specific risks.

In this paper, we consider AI systems that are IT systems. Several frameworks exist for cybersecurity risk management in IT systems and our goal is to simplify their adoption. We propose three blueprints that cover common deployments of AI systems and present a methodology to manage the risks based on these blueprints. Our approach is compatible with existing methodologies and can be either integrated with existing information security management systems or be used to start a new one.

We take into account cybersecurity, regulatory and AI-specific risks. As the cybersecurity risks in AI systems are similar to those in other IT systems, we focus here on AI-specific and regulatory risks. Our goal with this paper is to encourage more organizations to achieve basic AI security and safety.

2. AI IN THE CONTEXT OF SYSTEMS

Artificial intelligence has a variety of definitions in regulation and standards. Given the rapid pace of AI technology development, some definitions have become obsolete and need to evolve. The challenge lies in striking a balance between specificity and flexibility, ensuring that definitions are robust enough to guide current applications and future innovations.

The ISO/IEC 22989:2022 standard defines an AI system as an engineered system generating an output for a given set of human-defined objectives [1]. The definition in the draft European AI Act (AIA) [2] is similar but adds an (extendable) list of techniques and stresses interaction with its environment. The Organization for Economic Co-operation and Development (OECD) definition is similar to both but considers inputs and the autonomy of AI systems [3]. The United States' AI Bill of Rights uses the broader term automated systems [4].

A. Stakeholders of AI Systems

The listed sources also describe AI stakeholders. The ISO/IEC 22989:2022 standard provides a hierarchy of roles relevant to AI systems (see Section 5.19 and Figure 2 in [1]). The standard differentiates between AI providers, producers, customers, partners, subjects and relevant authorities. Such a differentiation is very helpful in discussing AI system deployments and their cybersecurity.

At the same time, for regulatory and legal discussions, definitions in regulation are more relevant. For example, the AIA definitions are more tailored for expressing responsibilities and mandates [2].

B. Components of AI Systems

The ISO/IEC 22989:2022 standard provides a functional view of AI systems with input, model processing and outputs as the main concepts. For systems based on machine learning, the concepts of training data, machine learning and continuous learning are added [1]. AIA defines the various kinds of data used in AI systems and discusses models and types of systems.

3. INCLUDING AI SYSTEMS IN RISK MANAGEMENT

A. Risk Management Frameworks and Standards

Risk assessment guidelines are defined by the ISO 31000 risk management standard [5] and the National Institute of Standards and Technology (NIST) risk management framework (RMF) described in NIST SP 800-37 [6]. These are refined for information security by ISO/IEC 27005 [7] and for cybersecurity by the NIST cybersecurity framework (CSF) [8]. Furthermore, ISO/IEC 23984 [9] provides AI-specific guidance for risk management, as does the NIST AI RMF [10].

By design, these standards and frameworks accommodate a wide range of possible systems, making them too complex to deploy in small organizations, especially if the team has no previous experience in risk management.

Our methodology starts with a generic three-step risk management process: context establishment, risk assessment and risk mitigation. The scope is defined as IT systems extended to be AI systems, and it provides guidance on identifying the most critical AI risks requiring action. The methodology aligns with ISO 31000 and ISO/IEC 27005, but should also be adaptable to the NIST RMF and CSF. Thus, organizations that later plan to adopt a standard risk management system can incorporate the work done using our methodology.

B. AI Considerations During Context Establishment

During context establishment, interested parties (including hidden ones) and relevant assets are documented, the risk appetite of an organization is defined, and risk owners are determined. The internal, national and regulatory requirements of interested parties are identified. Risk consequence, likelihood and acceptance criteria are determined, and a risk management approach is chosen. In this step, the party building an AI system needs to identify:

1) the data subjects or data owners on whose data the models have been trained;
2) the party who trained the model;
3) the party who is running the service; and
4) the service user.

The legal rights, obligations and motives of all parties must be taken into account. Each stakeholder can bring new applicable standards and regulations that need to be considered. Stakeholders can be considered as part of the organization or not.

The organization needs to identify where different types of data (models and training, input and output data) and software components (training, inference and data ingestion systems) originate and what the data flow is among the different components. Some risks may result from engaging with certain kinds of data or systems, so this mapping is a prerequisite.

Table I shows a simple way to map the relations between stakeholders and AI system components. Such visibility tables give a visual overview of the components to which each stakeholder has access. In this simple example, we have three stakeholders: the end user, the service provider (providing the AI front-end), and the AI application programming interface (API) provider (training and providing the model). All the stakeholders can see the end users' input data and the model output. The service provider and the AI API provider have access to the service provider's business data. Only the AI API provider can see the details of the model.

TABLE I: EXAMPLE VISIBILITY TABLE

	User input	Service provider data	AI model	Output
End user	X			X
Service provider	X	X		X
AI API provider	X	X	X	X

C. AI Considerations During Risk Assessment

Risk is often expressed in terms of the likelihood of a threat materializing and the severity of the consequences. The risk assessment phase roughly consists of risk identification, risk analysis and risk evaluation. During risk identification, the risks are found, recognized and described. The risk owners are also defined for each identified risk. During analysis, the causes, sources and likelihood of each risk, and the likelihood and severity of the consequences are determined. During evaluation, the results of the analysis steps are compared with risk criteria, prioritized and considered for risk treatment.

AI risk assessment builds on the context identified in the previous step. For each AI system component, we assess the risk in the context of the related stakeholders. The identification of this relationship is straightforward based on the visibility table compiled in the context establishment phase. For each identified stakeholder–component pair, we consider risks from three categories – cybersecurity, regulatory and AI-specific risks. Cybersecurity risks focus on the confidentiality, integrity and availability of AI system components (software, data and services). Regulatory risks deal with legal obligations that apply to stakeholders operating AI systems (for AI-specific regulations) or their components (e.g., regulations on personal data, copyrighted data or critical infrastructure). Finally, we define AI-specific risks as risks connected to the specificity of the algorithms, the impact of AI systems on our society and the ethical aspects of deploying AI.

Table II gives examples of defining a risk through vulnerabilities and threats. For each threat, the organization must determine the likelihood of it materializing and the severity of the consequences for its environment. The likelihood and severity of the same event can vary for different organizations.

In addition, it can be helpful to compare the risks of different deployments to decide on a solution for an organization. For instance, while a cloud provider may offer a wider range of security controls than a small organization can deploy by itself,

making an organization dependent on the cloud creates availability risks, should the connection to the cloud be lost.

TABLE II: EXAMPLE VULNERABILITY AND THREAT TABLE

Object	Risk category	Vulnerability	Threat
Output	AI risk	Biased or damaged model	End user will get an output that will direct them to act in a damaging way
Input	Regulatory risk	Insufficient legal basis for personal data processing	Service provider faces legal action over infringement of data protection regulations
Language model	Cybersecurity risk	Faulty identity management	AI API provider loses access to their infrastructure, stopping inference services

D. AI Considerations During Risk Treatment

There are several ways of treating risks, such as risk avoidance, risk modification, risk retention and risk sharing. The method is chosen based on the outcomes of the risk assessment process. Based on the prioritized list of risks sent for treatment, a set of necessary cybersecurity, AI security and regulatory controls will be determined so that the results will meet the organization's risk acceptance criteria.

The risk treatment of AI systems does not have any special steps. The controls for AI-specific risks can be different from regular cybersecurity controls, but the treatment is generally still done in the same way.

4. CYBERSECURITY AND REGULATORY RISKS AND CONTROLS

A. Cybersecurity Risks and Controls

As AI systems are cyberphysical systems, most standard cybersecurity controls apply. All stakeholders and components of AI systems can be considered stakeholders and assets in IT systems. A catalogue of cybersecurity controls can be found, for instance, in ISO/IEC 27002 [11] and NIST SP 800-53 [12].

B. Regulatory Risks and Controls

The legal landscape related to AI systems is developing rapidly. Regulations are being developed in the United States [13] and China [14]. In 2023, the European Union (EU) reached an agreement on the structure for a legal framework for AI. The regulation builds on a risk-based approach and distinguishes four types of AI systems: prohibited

AI (systems that manipulate user vulnerabilities, e.g., social scoring AI systems), high-risk AI, limited-risk AI, and minimal-risk AI [2], [15]. Specific transparency and disclosure requirements are provided for general-purpose AI systems [15]. Exceptions apply to research and development, open-source, national security, and military use [2]. A proposal for a directive on adapting non-contractual civil liability rules to AI is awaiting agreement [16].

In addition to legislation specifically regulating AI systems, there are also norms concerning product safety [17], [18], data protection [19]–[22], intellectual property [23], [24] and cybersecurity [25], [26], that must be followed. Sector-specific norms (e.g., in financial services or healthcare) and legal requirements in individual states will also apply. Finally, ethical principles [27] have to be followed. Together with the agreements between parties, they form the legal framework in which the AI system operates.[1]

In the case of legal aspects, especially in terms of liability, the role of the person in the AI system must also be taken into account. For example, duties and responsibilities to ensure compliance with the EU AI regulation vary by their role [2]. When processing personal data, data protection roles must also be taken into account [19], [28].

Non-compliance with legal requirements may lead to sanctions, including fines or suspension of operations until deficiencies are eliminated, as well as reputational damage and a decrease in system users. Litigation may also ensue if the rights of individuals are violated.

Control measures include the following:

1) Before starting the activity, prepare a legal scoping report to understand the legal framework in which you are operating. Map all applicable legislation, agreements and terms of service provisions, and keep the document up to date.
2) Some AI systems may need *ex-ante* conformity assessments and risk assessments [2]. A data protection impact assessment should be completed before processing personal data in an AI system (GDPR Art. 35; Directive (EU) 2016/680 Art. 27).
3) Ensure that you have relevant agreements, consents and licences for processing data (whether personal, copyrighted or other) throughout the life cycle of the AI system.
4) Implement organizational and technical measures to ensure both physical and digital security of the AI system and data throughout the entire AI system life cycle. Use appropriate privacy-enhancing technologies [29], [30].

1 It is crucial to familiarize oneself with all legal and contractual requirements to ensure the legality of all activities. The provided list of legislation is not exhaustive, so the relevant applicable legal framework must be assessed in each specific case.

5) Understand how the AI system works (human oversight), ensure system reliability and accuracy, apply a risk management system and best data governance practices through the system life cycle, prepare technical documentation and keep it up to date, and provide appropriate instructions and explanations to system users.

5. AI-SPECIFIC RISKS AND CONTROLS

A. Attacks Against AI Systems

AI systems make decisions based on data. These decisions can be critically important, can be based on sensitive data (e.g., in healthcare), or might have to be made in a split second and therefore lack human oversight (e.g., in self-driving cars or drones). These peculiarities of AI systems imply that the additional consideration of AI-specific threats is necessary for a complete risk analysis. The following is based on the German Federal Office for Information Security's 'AI security concerns in a nutshell' [31] as well as the Open Worldwide Application Security Project (OWASP) Foundation's 'OWASP Top 10 for LLM Applications' [32].

Evasion attacks are attacks where the attacker attempts to manipulate the model to return unexpected, incorrect or malicious outputs. For example, prompt injection is an evasion attack against a large language model (LLM), in which the attacker attempts to obtain an unauthorized output or have the model perform unauthorized actions by carefully constructing a prompt [33]. This prompt could be constructed with natural language, or it could utilize techniques such as adversarial suffixing [34]. Similarly, image classification models may be vulnerable to adversarial examples, where the model is manipulated into predicting an incorrect class through slight perturbations of the input [35]. Vulnerabilities related to evasion attacks can carry over in the case of transfer learning [36].

Information extraction attacks are attacks in which the attacker attempts to learn or reconstruct sensitive information such as model weights or training data. In an attribute inference attack, the attacker attempts to infer a sensitive attribute about an identity present in the training data by comparing the statistical relationships between features observed in model outputs with those that have been observed in the real world. Similarly, in a membership inference attack, the attacker tries to gain information about an identity's presence in the training data [37], [38]. In the case of model theft, the attacker attempts to construct a shadow model by feeding it training data gathered from model outputs. Another type of information extraction attack is model inversion, where the attacker attempts to reconstruct elements from the model's training data based on its outputs [39], [40].

Poisoning and backdoor attacks are aimed at training data. By altering the training data of an image recognition or a text-to-image model to associate certain inputs with a particular or a random incorrect label, the model could be steered to misclassify when detecting a particular object or its performance could be degraded in general [41], [42]. A backdoor attack is a more sophisticated form of data poisoning, where the labels are manipulated only when a particular trigger is present in an input.

Denial-of-service attacks, which are a type of availability attack, have some peculiarities in the context of AI applications. In autoregressive LLMs, for example, the processing power required to respond to a query depends on the length and content of the query. Lacking input validation or output length limits, an LLM could be made to generate very long outputs using its maximum context length, using up computational and memory resources and degrading performance for other users.

B. AI-Specific Risks

The adoption of AI for social, governance and industrial purposes as well as increasing reliance on AI have caused the emergence of previously unforeseen risks and ethical challenges. These risks can be broadly grouped into algorithmic and societal risks. Algorithmic risks are risks that arise from the technical aspects of an AI system and its application. For example, the output of a model might be biased, inaccurate or harmful, resulting from mistakes or attacks during model training. Societal risks emerge from the wider effects of AI on society and the unpredictability of future developments. The ethical challenges of AI adoption are broadly related to questions such as the choice of value models (e.g., utilitarian calculus in self-driving cars) as well as possible negative externalities of AI use.

Examples of AI-specific risks include:

1) Algorithmic risks: An AI system might fail to generalize on real-world data, underperform and give bad outputs. This is a particularly serious risk in critical applications such as healthcare, and is in turn amplified if the model is not explainable or lacks human oversight. In addition, a system might give outputs that are harmful or dangerous and, given bias in training data, discriminatory.
2) Societal risks: AI systems can expand the scope of human agency. This empowers users to not only do a lot of good but also to potentially cause considerable harm. As the speed of AI adoption and development outpaces regulatory efforts, there are significant risks of misuse. For example, malicious actors could use AI to aid them in developing weaponry. AI can be used to generate believable, high-quality disinformation, undermining

trust in online media. Autonomous AI agents could develop into an artificial superintelligence, which could pose an existential threat to humanity.

3) Ethical challenges: The use and possible misuse of AI raises a number of ethical questions. For example, a self-driving car might have to decide whether it is more appropriate to put its driver or a pedestrian at risk. Another example is exploitative or addictive applications, because empowering them with AI could increase the harm they pose to mentally and socially vulnerable individuals. There are also concerns such as ownership of AI-generated content, the implications of using AI in the justice system, and the ethics of job loss and other socioeconomic risks caused by AI versus the opportunity cost of inhibiting its adoption.

C. AI-Specific Controls

Evasion attacks can be mitigated by validating input prompts and outgoing requests, and monitoring model responses. Adversarial suffixing in LLMs as well as adversarial examples in image models can be mitigated by not releasing model weights publicly. If the model is connected to data sources or applications, it should never have more permissions than the user querying it and the model should be considered an untrusted user. To mitigate indirect prompt injections, in which LLMs accept compromised input from an external source, output received from external sources should be monitored and validated.

Information extraction attacks can be avoided to some degree by ensuring that training data does not contain sensitive personal data. In addition, an LLM application should never expose sensitive information in the pre-prompt. The model does not fundamentally distinguish between the prompt and the pre-prompt, so it should always be assumed that the user is capable of extracting it.

Data poisoning and backdoor attacks can be mitigated by scrutinizing the training data, applying quality criteria to filter it and validating its supply chain [43]. In addition, data poisoning can be detected at inference time by testing model performance for specific input categories.

To mitigate denial-of-service attacks, inputs should be validated, resource usage and API rate per user should be limited, and resource use should be monitored.

To ensure that an AI model performs consistently, performance should be monitored over time and across a diverse set of input categories. Similarly, to mitigate algorithmic risks related to bias or harmful outputs, safety metrics should be included in the monitoring process. Training data should be as diverse as possible. In addition,

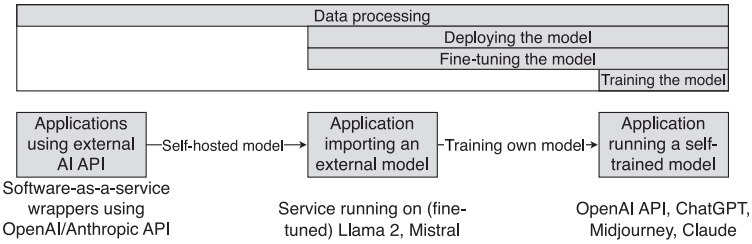
an effort should be made to make AI systems explainable, as explainability aids in interpreting monitored data, debugging the model and, thereby, achieving compliance.

6. GENERALIZED AI DEPLOYMENT BLUEPRINTS

A. Methodology

Architectural choices made in the design of an AI system significantly affect its risks. To simplify the risk analysis of new AI systems, we propose three generalized deployment blueprints (Figure 1). These blueprints differ in terms of the origin of the machine learning model, the party using the model on behalf of the service provider and data movement between parties. We consider AI systems with cloud services, as this is a common choice for AI systems where high computational performance is needed. However, the principles outlined here broadly apply to non-cloud applications as well. While these three deployment models do not cover every possible way to deploy an AI application, they can serve as guidance, helping application developers make sense of security and compliance risks.

FIGURE 1: OVERVIEW OF AI APPLICATION DEPLOYMENT MODELS WITH EXAMPLES



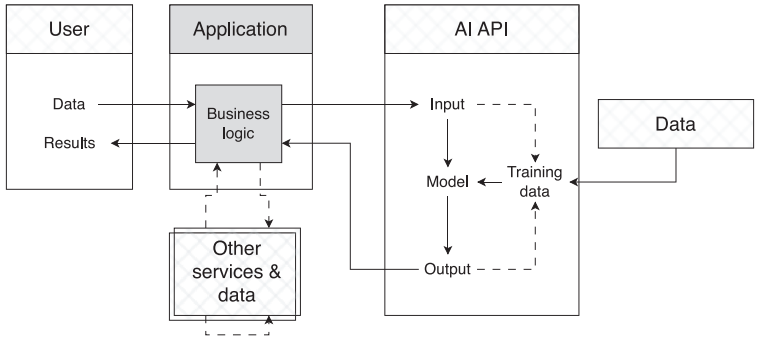
B. Systems Interfacing with an AI API

One way to deploy an AI system is by relying on a third-party API or AI-as-a-service in the business logic of the application (Figure 2). This permits the developer to leverage AI capabilities without having to deploy their own AI model for inference or having to train one on their own. In this deployment model, the system might also be interfacing with external data sources and services. The AI API might store the data it receives from the service to train its own models. This depends on the terms of service.

In this scenario, the AI model is external (not provided by the service provider), as is the training data. If the applications process user data, then user data moves to the application service, and from there to the AI API. The output is returned to the application service and then passed to the user after intermediate processing. It is

possible that user data is stored by both the service provider and the AI API provider. It is important to consider that relying on an external API externalizes availability risks. In addition, not all AI API providers support adapting (e.g., fine-tuning) the model to the service providers' needs.

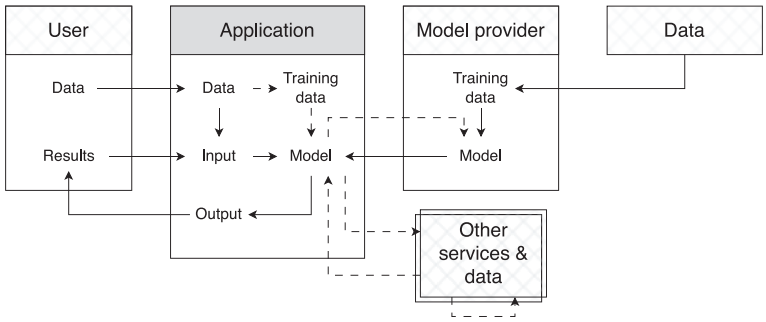
FIGURE 2: AI SYSTEMS RELYING ON AN EXTERNAL AI API



C. Systems Using a Self-Hosted External AI Model

By self-hosting a pre-trained model obtained from a model provider and (if necessary) adapting and fine-tuning it, some control over the deployment is regained (Figure 3). However, this deployment model introduces new challenges: the AI system owner now has to procure appropriate hardware and be more responsible for the model outputs, security and safety properties. In the case of fine-tuning, the AI system provider has to manage and curate the training data. In this scenario, user data is only stored by the service provider, unless third-party processors (e.g., cloud) are integrated into the system.

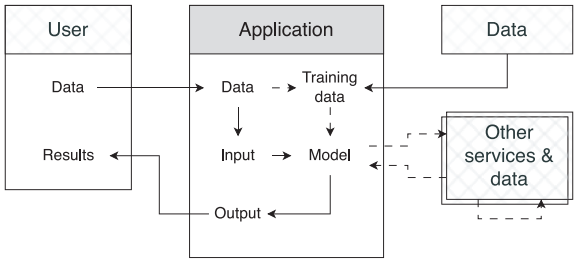
FIGURE 3: AI SYSTEMS IMPORTING AND SELF-HOSTING AN EXTERNAL MODEL



D. Systems Training Their Own Models

If an AI system is deployed using only self-trained and self-hosted AI models, risks related to data processing are internalized, as are those related to model performance and data quality (Figure 4). This deployment model is also used by dedicated AI technology providers who have the resources to deploy everything in-house, as well as by parties deploying the simplest, least computation-intensive AI applications, which do not require specialized hardware to train or host. Here, most data can be processed by the service provider.

FIGURE 4: AI SYSTEMS SELF-HOSTING A SELF-TRAINED MODEL



7. RISKS FOR AI SYSTEMS BASED ON THE BLUEPRINTS

Each of the presented three blueprints considers the system from the service provider's standpoint. Thus, the scope of the AI system under the control of the service provider increases. Service providers using an AI API are indirectly affected by risks related to inference or training. However, choices made by the stakeholders operating the inference and training components affect the service provider.

We provide a list of key risks to consider for service providers operating each blueprint. Once a service provider has identified that it is building an AI system that matches one of the three blueprints described in the previous section, it can make use of Table III to identify the key risks affecting that design.

The risks are grouped according to the three common stages in an AI workflow. The service making use of AI needs to process input data and output data using certain business logic on some infrastructure. This stage is shared by all blueprints. If the service provider opts to run inference with the AI model itself in order to limit data transfers to third parties, certain risks may be reduced, but the model needs to be

selected carefully. Finally, if the service provider trains the model itself, it needs to consider the risks related to processing training data and model quality. In the other blueprints, the service provider can try to make the AI API or AI model provider contractually liable for these concerns.

It must be noted that the risks listed in the table are the ones specific to AI systems. In each blueprint, common cybersecurity risks need to be assessed for each component for which the service provider is responsible. Also, each territory may have applicable local legislation or standards regulating the use of AI systems. These may add additional risks not included in the table.

We foresee that a user of this methodology will benefit from supportive tools (forms, tables, figures and worksheets) that help organize the information needed to follow the methodology. We have designed the first versions of such forms and the following guidance. These have not been included in this paper due to their size and are available in a separate report [44].

After performing an initial risk assessment using this table and picking the relevant mitigations, the service provider can set out to perform a full risk assessment according to a framework of their choice. The risk management conducted according to the blueprints in this paper will contribute to that analysis and ensure that it starts from a strong basis.

TABLE III: KEY RISKS FOR THE THREE GENERALIZED BLUEPRINTS

		External AI API	Self-hosted AI model	Self-trained AI model
Service provision	**Cybersecurity risks**	Availability of the AI API is not under the control of the service provider.	*Standard risks apply.*	*Standard risks apply.*
	Regulatory risk	Service provider does not have rights to process data, e.g., for transfer to the AI API provider, cloud or across borders.	Service provider does not have rights to process data, e.g., for transfer to cloud or across borders.	Service provider does not have rights to process data, e.g., for transfer to cloud or across borders.
	AI-specific risks	AI API uses a model that produces outputs that are unsafe or leak data.	*See inference risks below.*	*See inference risks below.*
Inference	**Cybersecurity risks**	*Not in scope of the service provider.*	Infrastructure used for AI inference does not perform well enough. Model provider does not provide updates.	Infrastructure used for inference does not perform well enough.
	Regulatory risks	*Not in scope of the service provider.*	Model contains data for which service provider does not have processing rights. Service provider does not have rights to process fine-tuning data.	*See training risks below.*
	AI-specific risks	*Not in scope of the service provider.*	AI model produces outputs that are unsafe or leak data. Data and tools used for fine-tuning reduce model quality.	*See training risks below.*
Training	**Cybersecurity risks**	*Not in scope of the service provider.*	*Not in scope of the service provider.*	Infrastructure used for model training does not perform well enough.
	Regulatory risks	*Not in scope of the service provider.*	*Not in scope of the service provider.*	Service provider does not have rights to process training data.
	AI-specific risks	*Not in scope of the service provider.*	*Not in scope of the service provider.*	AI model produces outputs that are unsafe or leak data. Data and tools used for fine-tuning reduce model quality.

8. FUTURE WORK

Further research includes comparing the proposed methodology with other lightweight risk management methodologies to quantitatively measure the effort needed for initial application and later maturation to a full quality management system. This will include a qualitative evaluation combining analysis of interview results with the evaluation of self-made assessments by professional risk managers.

A full report that details the background material and proposed methodology and provides supporting worksheets and a user's guide has been published by the Estonian Information System Authority [44]. Thus, we expect the methodology to find real-world use in both the public and private sector, providing opportunities to continue the suggested research.

REFERENCES

[1] *Artificial Intelligence – Artificial Intelligence Concepts and Terminology*. Standard ISO/IEC 22989:2022, ISO, 2022.

[2] 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act)', 2021/0106(COD), European Commission, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206

[3] 'Recommendation of the Council on artificial intelligence. Amended on: 08/11/2023', *OECD Legal Instruments*, OECD, 2023.

[4] White House Office of Science and Technology Policy, 'Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People', The White House, 2023.

[5] *Risk Management – Guidelines*. Standard ISO 31000:2018, ISO, 2018.

[6] *Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy*. Standard NIST SP 800-37 Rev. 2, US NIST, 2018.

[7] *Information Security, Cybersecurity and Privacy Protection – Guidance on Managing Information Security Risks*. Standard ISO/IEC 27005:2022, ISO, 2022.

[8] NIST Cybersecurity Framework 1.1. NIST, 2018.

[9] *Artificial Intelligence – Guidance on Risk Management*. Standard ISO/IEC 23984:2023, ISO, 2023.

[10] NIST AI Risk Management Framework 1.0. NIST, 2023.

[11] *Information Security, Cybersecurity and Privacy Protection – Information Security Controls*. Standard ISO/IEC 27002:2022, ISO, 2022.

[12] *Security and Privacy Controls for Information Systems and Organizations*. Standard NIST SP 800-53 Rev. 5. NIST, 2020.

[13] 'Fact sheet: President Biden issues executive order on safe, secure, and trustworthy artificial intelligence'. The White House. 30 Oct. 2023. [Online]. Available: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artif

[14] 'Interim Measures for the Management of Generative Artificial Intelligence (AI) Services'. Cyberspace Administration of China. 13 Jul. 2023. [Online]. Available: http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

[15] 'AI Act: Deal on comprehensive rules for trustworthy AI'. European Parliament. 9 Dec. 2023. [Online]. Available: https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai

[16] 'Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)', 2022/0303(COD), European Commission, 2022. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496

[17] Regulation of the European Parliament and of the Council of 10 May 2023 on General Product Safety, *Official Journal*, L 135, 23 May 2023, pp. 1–51. [Online]. Available: http://data.europa.eu/eli/reg/2023/988/oj
[18] 'Proposal for a Directive of the European Parliament and of the Council on liability for defective products', 2022/0302(COD), European Commission, 2022. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52022PC0495
[19] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, (General Data Protection Regulation), *Official Journal*, L 119, 4 May 2016, pp. 1–88. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj
[20] 'Proposal for a Regulation of the European Parliament and of the Council laying down additional procedural rules relating to the enforcement of Regulation (EU) 2016/679)', 2023/0202(COD), European Commission, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52023PC0348
[21] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016, *Official Journal*, L 119, 4 May 2016, pp. 89–131. [Online]. Available: http://data.europa.eu/eli/dir/2016/680/oj
[22] Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018, *Official Journal*, L 295, 21 Nov. 2018, pp. 39–98. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018R1725
[23] Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979). [Online]. Available: https://www.wipo.int/wipolex/en/text/283693
[24] WIPO Copyright Treaty. [Online]. Available: https://www.wipo.int/wipolex/en/text/295157
[25] Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022, *Official Journal,* L 333, 27 Dec. 2022, pp. 80–152. [Online]. Available: http://data.europa.eu/eli/dir/2022/2555/oj
[26] 'Proposal for a Regulation of the European Parliament and of the Council on horizontal cybersecurity requirements for products with digital elements and amending Regulation (EU) 2019/1020', European Commission, 2022. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52022PC0454
[27] European Parliament Resolution of 20 October 2020 with Recommendations to the Commission on a Framework of Ethical Aspects of Artificial Intelligence, Robotics and Related Technologies (2020/2012(INL)), *Official Journal*, C 404, 6 October 2021, pp. 63–106. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020IP0275
[28] 'Guidelines 07/2020 on the concepts of controller and processor in the GDPR', European Data Protection Board, 7 Jul. 2021. [Online]. Available: https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-072020-concepts-controller-and-processor-gdpr_en
[29] D. Bogdanov, E. Brito, P. Etti, L. Kamm, P. Laud, T. Mällo, A. Ostrak, K. Sein, R. Talviste, and M. Toomsalu. 'Privacy enhancing technology concept', (in Estonian), Cybernetica, Estonian Ministry of Economic Affairs and Communications, Tallinn, Estonia, 31 Mar. 2023. [Online]. Available: https://www.kratid.ee/_files/ugd/980182_f1288bebbb57466ead0241748d49d8ec.pdf
[30] D. Bogdanov, E. Brito, P. Etti, L. Kamm, P. Laud, T. Mällo, A. Ostrak, K. Sein, R. Talviste, and M. Toomsalu. 'Roadmap for deploying privacy enhancing technologies in Estonia', (in Estonian), Cybernetica, Estonian Ministry of Economic Affairs and Communications, Tallinn, Estonia, 31 Mar. 2023. [Online]. Available: https://www.kratid.ee/_files/ugd/980182_64478f7163b74f299f5879b6eea856af.pdf
[31] 'AI security concerns in a nutshell – practical AI-security guide', Federal Office for Information Security, Bonn, Germany, 2023.
[32] 'OWASP top 10 for large language model applications. Version 1.1', OWASP Foundation, 2023.
[33] Y. Liu, G. Deng, Y. Li, K. Wang, T. Zhang, Y. Liu, H. Wang, Y. Zheng, and Y. Liu. 'Prompt injection attack against LLM-integrated applications', 2024, *arXiv.2306.05499*.
[34] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Zico Kolter, and M. Fredrikson, 'Universal and transferable adversarial attacks on aligned language models', 2023, *arXiv.2307.15043*.
[35] I. J. Goodfellow, J. Shlens, and C. Szegedy, 'Explaining and harnessing adversarial examples', 2014, *arXiv.1412.6572*.
[36] J. Lin, L. Dang, M. Rahouti, and K. Xiong, 'ML attack models: Adversarial Attacks and Data poisoning attacks', 2021, *arXiv.2112.02797*.
[37] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, 'Membership inference attacks against machine learning models', *IEEE Symposium on Security and Privacy (SP)*, USA, pp. 3–18, 2017, doi: 10.1109/SP.2017.41.
[38] B. van Breugel, H. Sun, Z. Qian, and M. van der Schaar, 'Membership inference attacks against synthetic data through overfitting detection', 2023, *arXiv.2302.12580*.

[39] N.-B. Nguyen, K. Chandrasegaran, M. Abdollahzadeh, and N.-M. Cheung, 'Re-thinking model inversion attacks against deep neural networks', *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Canada, pp. 16384–16393, 2023, doi: 10.1109/CVPR52729.2023.01572.
[40] K.-C. Wang, Y. Fu, K. Li, A. Khisti, R. Zemel, and A. Makhzani, 'Variational model inversion attacks', 2022, *arXiv.2201.10787*.
[41] S. Shan, W. Ding, J. Passananti, H. Zheng, and B. Y. Zhao, 'Prompt-specific poisoning attacks on text-to-image generative models', 2024, *arXiv.2310.13828*.
[42] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Madry, B. Li, and T. Goldstein. 'Dataset security for machine learning: Data poisoning, Backdoor Attacks, and Defenses', *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, no. 2, pp. 1563–1580, 2023, doi: 10.1109/TPAMI.2022.3162397.
[43] I. Brown. 'Expert explainer: Allocating accountability in AI supply chains'. 2023. [Online]. Available: https://www.adalovelaceinstitute.org/resource/ai-supply-chains
[44] D. Bogdanov, P. Etti, L. Kamm, A. Ostrak, F. Stomakhin, M. Toomsalu, S.-M. Valdma, and A. Veldre 'A study of the risks and controls for artificial intelligence and machine learning technologies 1.0', (in Estonian), Cybernetica, Estonian Information System Authority, 27 Feb 2024. [Online]. Available: https://www.ria.ee/sites/default/files/documents/2024-03/Tehisintellekti-masinoppe-tehnoloogia-riskide-uuring-2024.pdf

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

ERS0: Enhancing Military Cybersecurity with AI-Driven SBOM for Firmware Vulnerability Detection and Asset Management

Max Beninger
Research Assistant
School of Computing
Queen's University
Kingston, ON, Canada
max.beninger@queensu.ca

Philippe Charland
Defence Scientist
Mission Critical Cyber Security Section
Defence Research and Development Canada
Quebec, QC, Canada
philippe.charland@drdc-rddc.gc.ca

Steven H. H. Ding
Assistant Professor
School of Information Studies
McGill University
Montreal, QC, Canada
steven.h.ding@mcgill.ca

Benjamin C. M. Fung
Professor
School of Information Studies
McGill University
Montreal, QC, Canada
ben.fung@mcgill.ca

Abstract: Firmware vulnerability detection and asset management through a software bill of material (SBOM) approach is integral to defensive military operations. SBOMs provide a comprehensive list of software components, enabling military organizations to identify vulnerabilities within critical systems, including those controlling various functions in military platforms, as well as in operational technologies and Internet of Things devices. This proactive approach is essential for supply chain security, ensuring that software components are sourced from trusted suppliers and have not been tampered with during production, distribution, or through updates. It is a key element of defense strategies, allowing for rapid assessment, response, and mitigation of vulnerabilities, ultimately safeguarding military capabilities and information from cyber threats.

In this paper, we propose ERS0, an SBOM system, driven by artificial intelligence (AI), for detecting firmware vulnerabilities and managing firmware assets. We harness

the power of pre-trained large-scale language models to effectively address a wide array of string patterns, extending our coverage to thousands of third-party library patterns. Furthermore, we employ AI-powered code clone search models, enabling a more granular and precise search for vulnerabilities at the binary level, reducing our dependence on string analysis only. Additionally, our AI models extract high-level behavioral functionalities in firmware, such as communication and encryption, allowing us to quantitatively define the behavioral scope of firmware. In preliminary comparative assessments against open-source alternatives, our solution has demonstrated better SBOM coverage, accuracy in vulnerability identification, and a wider array of features.

Keywords: *vulnerability detection, firmware analysis, firmware management, artificial intelligence*

1. INTRODUCTION

The implementation of a software bill of materials (SBOM) practice in military operations is becoming increasingly critical, particularly in the context of supply chain security, asset management, and vulnerability management [1], [2]. Military operations typically rely on a complex network of platforms, operational technologies, and their software components that are often sourced from a myriad of suppliers, each with varying levels of trust and transparency. The complex nature of modern supply chains in software procurement and deployment inherently assumes a high level of trust in all participating suppliers. This trust-based approach, however, exposes these supply chains, particularly open-source software (OSS), to significant risks of supply chain attacks [3]. Such attacks can occur when a malicious actor infiltrates the supply chain at any point, potentially compromising the integrity and security of the software components being distributed or the pipelines through which the components are produced and integrated. This systematic weakness is especially concerning in environments where software plays a critical role in operational functionality and security, such as in military operations.

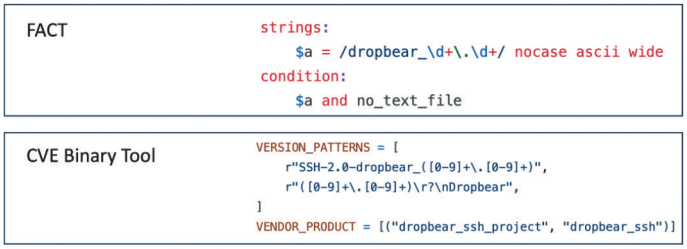
An SBOM serves as a strategic tool to mitigate these risks by introducing an element of transparency into the supply chain. By providing a comprehensive and detailed list of all software components used in a system, including their origins, versions, and dependencies, an SBOM makes it possible to scrutinize and validate each component's security and integrity. This level of transparency is crucial in identifying and addressing vulnerabilities that might otherwise go unnoticed in the complex

web of supply chain relationships. As a proactive approach, an SBOM is particularly valuable in preventing supply chain attacks, as it allows for the early detection of any anomalies or unauthorized alterations in the software components.

Obtaining a reliable SBOM is challenging, as relying on suppliers to provide this crucial information reintroduces the very trust issues SBOMs are meant to mitigate. Expecting suppliers to disclose the complete source code for all released firmware is also unrealistic, not only due to proprietary concerns but because the pipeline that transforms source code into final firmware can itself be compromised within the supply chain. Consequently, a shift-right approach is necessary, predicated on a zero-trust assumption toward suppliers. This approach seeks the development of specialized tools capable of directly extracting SBOMs from the already released or deployed firmware and software components. Such tools would independently analyze and generate a comprehensive list of components, bypassing the need to rely on supplier-provided information and ensuring a more accurate and secure assessment of the software's composition and potential vulnerabilities.

Existing solutions for SBOM generation primarily depend on manual processes, where specific string-matching patterns are crafted. These patterns are designed to detect various versions of strings linked to a particular product or open-source project. An example of this can be seen in Figure 1, which displays manually created regular expressions aimed at identifying different versions of the Dropbear library. This manual process is not only labor-intensive but also demands expertise in firmware analysis, given the extensive range of pattern variants necessary for effective matching. As a result, existing state-of-the-art solutions, such as the Firmware Analysis and Comparison Tool (FACT) [4] and the Common Vulnerabilities and Exposures (CVE) Binary Tool [5], are limited in their scope, only able to identify a few hundred specific products and components. This limitation underscores the need for more advanced and automated methods in SBOM generation. Moreover, these solutions often overlook code-level patterns, which are crucial as they contain the actual vulnerabilities. The failure to capture these patterns represents a significant gap in the effectiveness and thoroughness of current SBOM generation methods. Another inadequately addressed aspect is the characterization of firmware's high-level capabilities, such as encryption, communication, and I/O operations. These capabilities are crucial for military operations, as they define the operational scope, the potential attack vectors, the presence of a possible backdoor, and the integrity of the firmware overall. Therefore, there is a pressing need for a solution that not only identifies the components of an SBOM but also comprehensively understands and delineates the high-level functionalities and potential capabilities of the firmware.

FIGURE 1: EXAMPLES OF MANUALLY CREATED RULES FOR VERSION STRING MATCHING IN TWO DIFFERENT SBOMS TOOLS



To address these challenges, we propose a novel system, ERS0, that leverages machine learning to scale up and expand the SBOM creation process. Our methodology includes two key components: (1) character-level string similarity learning for precise version string and product detection, and (2) cross-architecture code clone search for OSS library and version identification. This approach considerably broadens the scope and accuracy of SBOM analysis. Furthermore, we aim to detect high-level capabilities over firmware images using static analysis. We have developed a generative large language model (LLM) that is up-trained to create capability identification rules based on the ATT&CK behavior catalog [6]. ERS0 not only enhances the accuracy of SBOM identification but also provides a comprehensive understanding of the firmware's functionalities and potential vulnerabilities, thereby substantially contributing to the security and robustness of military operations. Our contributions can be summarized as follows:

- We propose a learning-based version string-matching approach to address the scalability of the original manual process. Our proposed system, ERS0, is capable of accommodating over 1.4 million variant packages across a diverse range of products.
- We propose an efficient SBOM generation method based on cross-architecture assembly code clone search. This technique specifically targets the previously unaddressed gap in code-level analysis.
- We develop a generative LLM designed to create rules that effectively match string and code patterns correlated with the high-level behavioral capabilities of firmware.

This paper is organized as follows. Section 2 discusses the related tools in this domain. Section 3 describes the overall workflow of the ERS0 system and elaborates on each individual component. Section 4 demonstrates the effectiveness of our system, and Section 5 presents its interface. Finally, Section 6 concludes this paper.

2. RELATED WORKS

SBOM is becoming increasingly important for software security, especially for identifying components in software packages. SBOM generation is a relatively new area of research. Tools such as FACT, the CVE Binary Tool, and EMBArk are well known in this domain, but they have some scalability drawbacks. FACT helps to break down and examine firmware, which is important for generating SBOMs for embedded systems and Internet of Things devices. FACT is good at automatically finding and analyzing parts of firmware, helping to list all software parts needed for an SBOM. However, it requires manually created rules to identify these parts, which makes it less effective when there is a variety of different products. As more products with different types of firmware come out, updating these rules by hand can lead to missing or old information in SBOMs.

The CVE Binary Tool [5] scans software to find known vulnerabilities, adding to the SBOM by identifying potential security issues in the software components. It looks at software for versions that have known problems, which helps SBOMs show possible security risks. But, like FACT, the CVE Binary Tool also depends on rules that need to be manually generated. Keeping these rules up to date is labor-intensive, especially when software changes quickly and comes in many forms. This makes it hard for the tool to keep up and cover a wide range of software components accurately.

EMBArk [7] focuses on firmware security and is important for making SBOMs for embedded systems. It does a good job of automatically analyzing firmware and giving detailed information needed for SBOMs. But EMBArk also has the same problem as the other tools: It requires manually made rules to find parts and security issues. Writing these rules takes a lot of work and does not keep pace with the fast changes in software and the many different products available. This makes it less useful for generating SBOMs for a large number of products.

In short, while these tools are helpful for producing SBOMs and analyzing software security, their need for manually made rules makes it hard to use them for a wide range of products. This is a serious problem, because software is always changing and there are so many different types of software products available. We need ways to make and update SBOMs that are faster and can handle many different products at once.

3. SYSTEM DESIGN

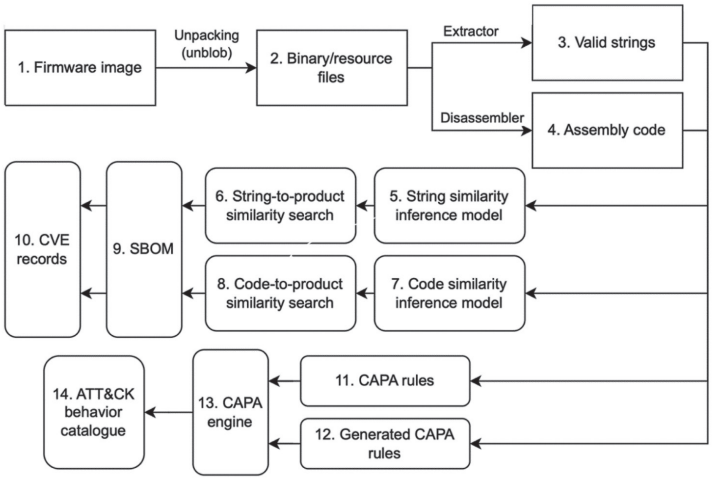
The automated SBOM generation process in ERS0 is a general workflow that starts with a firmware image and results in a thorough SBOM, complete with security vulnerability assessments and behavior analysis. This detailed process integrates the use of an open-source unpacking tool, machine learning models, and both predefined and dynamically generated rules to produce an SBOM that is informative not only in terms of component listing but also in terms of security analysis. The workflow, shown in Figure 2, is as follows:

- **Unpacking, extraction, and disassembly (Steps 1–4):** The analysis begins with a firmware image that will be scrutinized for its software contents. Utilizing unblob [8], an open-source tool that supports many image formats, the firmware is unpacked to isolate binary and resource files. After this, the extraction utility identifies valid strings from the binaries, which can be indicative of component names, versions, and other key metadata. Next, a disassembler abstraction module, compatible with both the Ghidra [9] and IDA Pro [10] disassemblers, converts the machine code of binary files into assembly code, which can be analyzed for richer insights.
- **String embedding and string-to-product search (Steps 5–6):** A machine learning model that transforms strings into an embedding space to identify their features and relationships to version strings, aiding in the identification of software components. An embedding space is a high-dimensional vector space where strings are represented as points or vectors. This representation facilitates the comparison of strings by measuring distances or angles between them, allowing for the identification of similar products based on their proximity in the space. Next, this module compares the resulting string embeddings to known product name embeddings to identify potential matches in software components.
- **Code embedding and code-to-product search (Steps 7–8):** Like the string model, this model transforms segments of assembly code into an embedding space to detect patterns and features that correspond to known software code components. The model is trained to match semantically similar assembly code across different platforms. This step involves comparing code embeddings with known product code embeddings to identify components within the firmware.
- **SBOM and CVE records matching (Steps 9–10):** All identified components, along with their versions and interrelations, are compiled into an SBOM. With the SBOM, ERS0 conducts a review of CVE records to determine if any of the identified components are associated with known vulnerabilities.

- **CAPA rules and generated CAPA rules (Steps 11–13):** ERS0 integrates the CAPA engine [11]. It is a part of the analysis system that applies predefined and dynamically generated rules to evaluate the capabilities and behaviors of software based on its code and metadata. It uses these rules to detect patterns that could signify potential security threats, vulnerabilities, or malicious activities within the software being analyzed. ERS0 uses predefined, manually created rules within the CAPA engine to abstract software capabilities from observed behaviors for threat analysis. In addition to the predefined rules, supplementary CAPA rules are generated by our language model to enhance the behavioral analysis.
- **ATT&CK behavior catalog (Step 14):** The system catalogs the identified behaviors according to the MITRE ATT&CK® framework [6], which serves as a global knowledge base of adversary tactics and techniques. More details are provided in Section 2.C.

Overall, the automated SBOM generation workflow facilitates not only the identification and documentation of software components but also the assessment of their security risks, thereby offering a robust tool for software component transparency and risk management.

FIGURE 2: OVERALL WORKFLOW OF ERS0 SYSTEM FOR FIRMWARE SBOM AND CAPABILITY ANALYSIS



A. Similarity Learning for Version String Matching

Instead of employing manual rule creation, we propose the use of a machine learning model to efficiently capture string patterns associated with various products on a large

scale. This machine-learning model is designed to convert a given input string into a numeric vector embedded within a high-dimensional space. Specifically, the model aims to ensure that the vector representation of a version string closely aligns, in terms of angular similarity, with its corresponding product name. Conversely, an invalid version string or one belonging to a different product should exhibit dissimilarity with the product name in this embedded space. By adhering to this principle, the model can effectively memorize how different variations of version strings should relate to their associated product names.

For instance, let us take the product name "pdns" and its valid version string "[lua2backend]," which corresponds to the lua2 backend version 4.7.3. Here, the model is trained to yield a similarity value of 1, since "[lua2backend]" is a valid version string for pdns. However, it should produce a similarity value of 0 when presented with the OpenSSL product name. Invalid version strings are those that may appear to conform to the standard format of major, minor, and build versions, but do not accurately represent the product's version. For example, the string "IEEE 802.11" might seem like a valid version string with "802.11" as its version number. However, it pertains to the network communication protocol and should not be treated as an SBOM entry. Therefore, it should yield a similarity value of 0 when compared to all product names. The model is trained on a dataset that follows this format, comprising over 4.3 million unique valid version string patterns. The dataset comprises pre-compiled Linux libraries, excluding Windows binaries, to test cross-platform generalizability.

The model itself operates as a character-level language model. It accepts a raw string as input, applies character-level tokenization, converts the raw string into a sequence of characters, and encodes each character into a numeric vector (embedding) using multiple layers of transformer models. Specifically, we leverage the CANINE [12] pre-trained language model as the encoder. Additional details can be found in the original paper [12]. This model has been pre-trained on an extensive text corpus, encompassing various domains such as news postings, Wikipedia articles, and programming questions/answers. Leveraging this pre-trained language model facilitates semantic matching between product names and version strings. For example, "libcrypto" is a library within the "libssl" package, which is part of the OpenSSL project. The pre-trained language model, without further fine-tuning, can already establish a high degree of similarity between "libcrypto" and both "libssl" and "openssl."

As shown in Figure 3, the encoder model is up-trained following a Siamese architecture. A Siamese network is a type of neural network architecture used for learning similarity or dissimilarity between pairs of data points [13]. In our case, the goal is to measure the similarity between product names and version strings. Let us denote the input product name as P and the input version string as V. These inputs are

converted into numeric vectors using the CANINE pre-trained language model. Let $f(P)$ and $f(V)$ be the embeddings (numeric vectors) of P and V, respectively:

$$f(P) = \text{CANINE}(P)$$
$$f(V) = \text{CANINE}(V)$$

The cosine similarity between the embeddings $f(P)$ and $f(V)$ is calculated as:

$$\text{cosine\_similarity}(f(P), f(V)) = \frac{f(P) \cdot f(V)}{\| f(P) \| \cdot \| f(V) \|}$$

Now, the cosine loss function can be defined as follows:

$$\text{Cosine\_Loss}(f(P), f(V), \text{label}) = \begin{cases} 1 - \text{cosine\_similarity}(f(P), f(V)), & \text{if label} = 1 \\ \text{cosine\_similarity}(f(P), f(V)), & \text{if label} = 0 \end{cases}$$

We want to minimize the cosine similarity between valid pairs (product name and its corresponding version string) and maximize the cosine similarity between invalid pairs (product name and a version string of a different product). In practice, a label of 1 is assigned to denote a valid pair of strings, while a label of 0 is used for an invalid pair of strings.

During the deployment stage, as shown in Figure 4, the first step involves encoding all existing product names into vector representations. These vectors serve as a reference database against which incoming strings can be compared. The encoding is carried out by the above trained encoder, which transforms raw string data into a high-dimensional space where semantically similar terms are placed closer together.

FIGURE 3: A SIMILARITY-BASED MACHINE LEARNING MODEL FOR THE VERSION STRING-TO-PRODUCT NAME-MATCHING PROBLEM

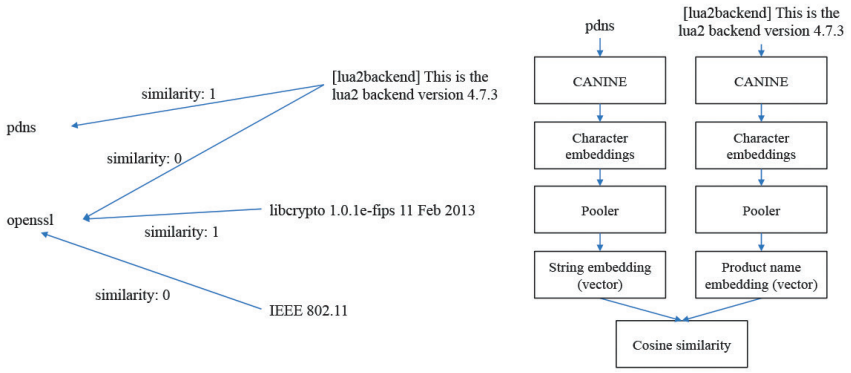
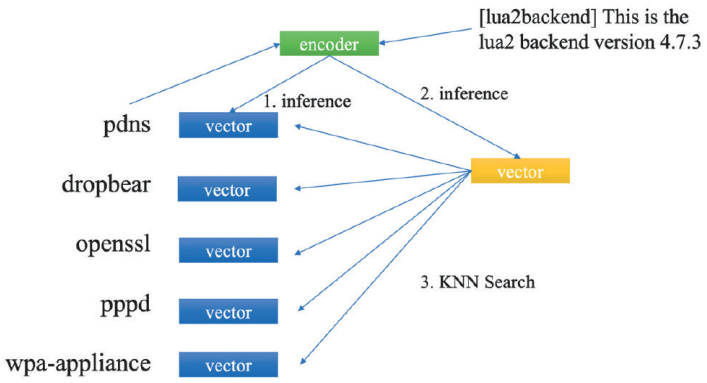


FIGURE 4: MATCHING AND SEARCHING PROCESS DURING THE DEPLOYMENT STAGE



When a new incoming string, which could be a potential version string, is received, the same encoder processes it to generate its vector representation. This vector is then compared against the pre-encoded vectors of product names, using a cosine similarity measure. The process searches for the vector among the pre-encoded product names that are most similar to the vector of the incoming string, limited by a predefined threshold. If a vector that exceeds this threshold is found, the incoming string is considered to be a match for the corresponding product name. This method allows for robust string matching, accommodating variations and minor discrepancies that often occur in real-world data.

B. Code Clone Search for Library Matching

This module functions in a way that is similar to the string-matching module, but it has a different data target. Its primary goal is to compare the assembly code of a given binary executable with the code of known product binaries. It utilizes a specifically trained model known as Pluvio [14] to identify similarities in assembly code across various computer architectures. This model goes beyond the traditional approach of matching version strings to product names. Instead, Pluvio is designed to detect assembly code that is similar in function and purpose across different types of platforms and against variations of the compilation toolchain. For example, a training sample involves the task of comparing two versions of a checksum algorithm CRC32 [15], one compiled for an ARM processor and the other for an AMD64 processor. The model's training on such examples enables it to recognize functions that have similar purposes or behaviors, even though they might be compiled in different ways.In the operational workflow, ERS0 first disassembles the binary executable. This task is carried out using a selected backend disassembler, as explained in the overall

workflow (referenced in Figure 2). The disassembler's job is to convert the machine code into assembly code, thereby revealing richer semantic information about the code's meaning and structure. Subsequently, the disassembler performs a control flow analysis. This analysis is crucial, as it segments the continuous stream of assembly code into individual and distinct assembly functions. Each binary executable is decomposed into a comprehensive list of assembly functions, making it easier to analyze and compare. When a new binary executable is extracted from a firmware, ERS0 compares it against a vast repository of known assembly functions from a wide range of products, extracted from open-source Linux and Windows packages. This repository is extensive and includes a diverse array of functions, providing a robust basis for comparison. By comparing the new executable to this repository, ERS0 can effectively identify any similarities or matches.

In the code matching algorithm (Algorithm 1), the process begins with Step 1, where the target binary is inputted as the subject of the search and analysis. Step 2 involves extracting all the assembly functions from the target binary, which requires disassembling the binary to understand its low-level code structure. In Step 3, each function *f* extracted from the target binary is iterated through for individual analysis. Step 4 uses the Pluvio model to search for clones of each function *f* in a comprehensive repository; these clones are variants of the function found in different binaries, and the search focuses on matches that meet a specified similarity threshold. Step 5 involves counting the origin of every clone identified in the previous step, thereby increasing a counter corresponding to its source binary or library; this step is crucial for tracking which binaries or libraries have functions most similar to those in the target binary. Step 6 involves selecting the top 10 binaries or libraries with the highest clone count, representing the binaries/libraries whose functions most frequently matched with those in the target binary. Finally, Step 7 refines this process by searching each function *f* from the target binary again, but only within the top 10 previously identified binaries/libraries, to ensure an accurate clone count. In Step 8, the algorithm checks the clone count threshold and picks the highest matched source binary or library, thus concluding the binary analysis and identification process.

Algorithm 1: Binary assembly function matching algorithm

Input: Target binary executable query Q, matching threshold θ, top k parameter k, match ratio threshold ρ
Output: Matched source binary/library ▷ Step 1: Input target binary
procedure MatchAssemblyFunctions(Q, θ, k, ρ)
    $F \leftarrow$ RetrieveAssemblyFunctions(Q) ▷ Step 2: Retrieve assembly functions
    $Counter \leftarrow$ InitializeCounter ▷ Initialize counter variable
    for each function $f \in F$ **do** ▷ Step 3
        $C_f \leftarrow$ SearchClonesAgainstCompleteRepository(f, θ, k) ▷ Step 4
        for each clone $c \in C_f$ **do** ▷ Step 5
            IncreaseCounter($Counter$, origin(c))
        end for
    end for
    $B \leftarrow$ GetTopBinariesWithHighestCount($Counter, k$) ▷ Step 6
    for each function $f \in F$ **do** ▷ Step 7
        for each binary $b \in B$ **do**
            SearchFunctionInLibraries(f, b)
        end for
    end for
    $S \leftarrow$ CheckCloneCountThreshold($Counter, \rho$) ▷ Step 8
    return PickHighestMatchedSource(S)
end procedure

C. Generative AI for Behavior Rule Generation

The final phase of the ERS0 project focuses on the development of behavior-matching rules. The objective is to leverage a large pre-trained natural language model, such as ChatGPT or Llama2, for generating YARA rules. YARA rules are essentially patterns or sets of conditions used to identify and classify binary executables [16]. Originally, YARA rules were used to classify and label malware, but their applications have since been extended beyond malware analysis into the domain of general binary analysis for threat hunting. CAPA is an open-source project leveraging YARA rules to identify high-level capabilities in binary executables. Figure 5 shows an example YARA rule. It tries to identify the "ws2_32. select" application programming interface (API) and label the binary with the capability to get socket status. The CAPA project and its YARA rules are crucial for identifying specific behaviors cataloged in ATT&CK, a comprehensive database of techniques employed by malicious entities. The example rule also defines the corresponding ATT&CK technique label T1016. T1016 in the MITRE ATT&CK framework refers to "System Network Configuration Discovery." This technique is part of the discovery tactic, where attackers seek to gather information about your network and systems, which can then be used to guide further actions, such as lateral movement through the network or understanding what defenses are in place.

FIGURE 5: EXAMPLE YARA RULE FROM THE CAPA OPEN-SOURCE PROJECT TO FIND CAPABILITIES PRESENTED IN A BINARY EXECUTABLE

```
rule:
  meta:
    name: get socket status
    namespace: communication/socket
    authors:
      - michael.hunhoff@mandiant.com
    scopes:
      static: function
      dynamic: call
    att&ck:
      - Discovery::System Network Configuration Discovery [T1016]
    mbc:
      - Communication::Socket Communication::Get Socket Status [C0001.012]
    examples:
      - 6A352C3E55E8AE5ED39DC1BE7FB964B1:0x1000C1F0
  features:
    - and:
      - api: ws2_32.select
```

However, the rule creation process relies on manual effort, requiring a security analyst with a strong background in understanding ATT&CK techniques, as well as strong familiarity with both low- and high-level programming APIs and libraries across many platforms. This poses a significant challenge, as the rule generation process must be scaled up to cover all ATT&CK techniques. We employ Retrieval Augmented Generation (RAG) prompt engineering to facilitate the rule creation process.

RAG combines language models with a retrieval system to generate text using external information [17]. After starting a prompt conversation, it first retrieves relevant information about the query to enrich the conversation context and then proceeds to the subsequent tasks provided in the prompt conversation. ERS0 uses the Atomic Red Team project [18] on GitHub as the external knowledge base. It contains both the technical description of each ATT&CK technique and several implementations of example attacks. This RAG design allows AI to provide up-to-date and domain-specific information dynamically. For each ATT&CK technique, ERS0 starts a prompt conversation with eight steps to create a YARA rule:

- **Step 1: Data provision.** In this step, the entire ATT&CK catalog, including technique descriptions and attack examples, is inputted into the language model from the Atomic Red Team project. This comprehensive data ensures that the model has a deep understanding of various attack techniques. This understanding is critical for accurate and effective rule generation.

- **Step 2: Technique identification.** This step involves requesting the model to provide a detailed description for a given ATT&CK technique ID. By understanding the specifics of the given technique, the model can tailor the generated rule to precisely match the behavior associated with that technique.
- **Step 3: Similarity analysis.** Here, the model is tasked with finding and listing the top five techniques most similar to a given ID based on their descriptions. This analysis helps enrich the context of the technique being targeted for rule generation and improves the accuracy of the rule being generated.
- **Step 4: System call identification.** This step instructs the model to identify typical Windows and Linux system calls that are related to the technique under consideration. Recognizing and incorporating relevant system calls is essential for pinpointing specific behaviors that need to be addressed by the rule.
- **Step 5: Advanced command analysis.** Advanced analysis involves having the model identify less common and undocumented Windows kernel APIs [19], if applicable, linked to the technique. These obscure APIs are often exploited in sophisticated attacks and including them in the rule ensures that such tactics are not overlooked.
- **Step 6: Cross-platform analysis.** Extending the analysis to include identifying relevant MacOS, Java, and Python APIs for the technique ensures that the generated rule is comprehensive and effective across different operating systems. This step broadens the scope of rule applicability.
- **Step 7: Constant value identification.** In this step, the model is asked to identify all potential constant values, such as hexadecimal numbers, floating-point numbers, decimal numbers, or string values, used by each of the APIs identified above. Constants often serve as key indicators in malicious operations, making their identification crucial for rule accuracy.
- **Step 8: YARA rule generation.** Finally, based on the information gathered in previous steps, the model is instructed to generate a YARA rule specifically tailored to detect the behavior associated with the targeted technique. This rule serves as a practical tool for identifying the presence of a specific malicious technique within a system.

By automating these steps, ERS0 overcomes the challenges of manual rule creation, enabling the generation of comprehensive and effective detection rules at a much larger scale. This method is not only efficient but also ensures that the rules are up-to-date and relevant across various systems.

4. EXPERIMENTAL EVALUATION OF SBOM MATCHING

We developed a benchmark dataset for SBOM generation, essential for firmware analysis to evaluate a tool's coverage of libraries and potential vulnerabilities. We assessed the performance of three tools: FACT, EMBArk, and the CVE Binary Tool, each in conjunction with ERS0. Our dataset was compiled by aggregating all packages from two sources: the Debian package repository for Linux pre-compiled packages and the Conan open-source package repository for MS Windows dynamic-link libraries (DLLs). This resulted in a total of 395,933 packages and 1,583,732 binary executables, each annotated with product and version information.

For our evaluation, we tested each tool individually along with ERS0. This approach was chosen because ERS0 is designed to be compatible with all the packages in our dataset, whereas each of the other tools has varying levels of support for different packages. We specifically assessed a tool on a package only if it had a predefined rule that corresponds to that package.

The results depicted in Table I provide a comparative analysis of three tools—FACT, EMBArk, and the CVE Binary Tool—in their performance of matching binaries to applicable packages, alongside the performance of ERS0. We assess tools based on their accuracy in matching binary executables, such as libcrypto.dll, to their correct vendor (e.g., OpenSSL) and version (e.g., 1.1.1). A correct match scores 1.

TABLE I: EXPERIMENTATION RESULTS

	FACT	EMBArk	CVE Binary Tool
Total number of applicable packages	42,838	78,854	102,571
Total number of applicable binaries	54,394	218,433	305,558
Number of binaries matched by the tool	129	1,075	46,495
Number of binaries matched by ERS0 and match percentage of Code-To-Product method	45,690 (14.3%)	163,824 (24.5%)	269,432 (16.1%)

FACT had a total of 42,838 applicable packages with 54,394 binaries. Of these, FACT matched 129 binaries, while ERS0 identified significantly more, with 45,690 matches. EMBArk, with 78,854 applicable packages and 218,433 binaries, matched 1,075 binaries. ERS0 again showed a higher matching count with 163,824 binaries. The CVE Binary Tool had the highest number of applicable packages at 102,571 and the highest number of applicable binaries at 305,558, with the tool itself matching 46,495

binaries. In this case as well, ERS0 demonstrated a more extensive matching capability, identifying 269,432 binaries. The effectiveness of code-to-product matching depends on the disassembler's performance and availability, but this limitation is mitigated by the string-to-product matching feature.

These results suggest that while each tool has its own capacity to match binaries, ERS0 shows a more robust performance in identifying binary matches across all tools. The high number of matches by ERS0 could indicate its higher matching coverage capability. However, the lower match counts for the individual tools do not necessarily imply poor performance; they may be highly specialized or conservative in their matching to ensure high accuracy. It is also important to consider the context and specific use cases in which one tool might perform better than the others despite the lower numbers, such as in scenarios requiring specific package support or higher precision.

5. SYSTEM INTERFACE

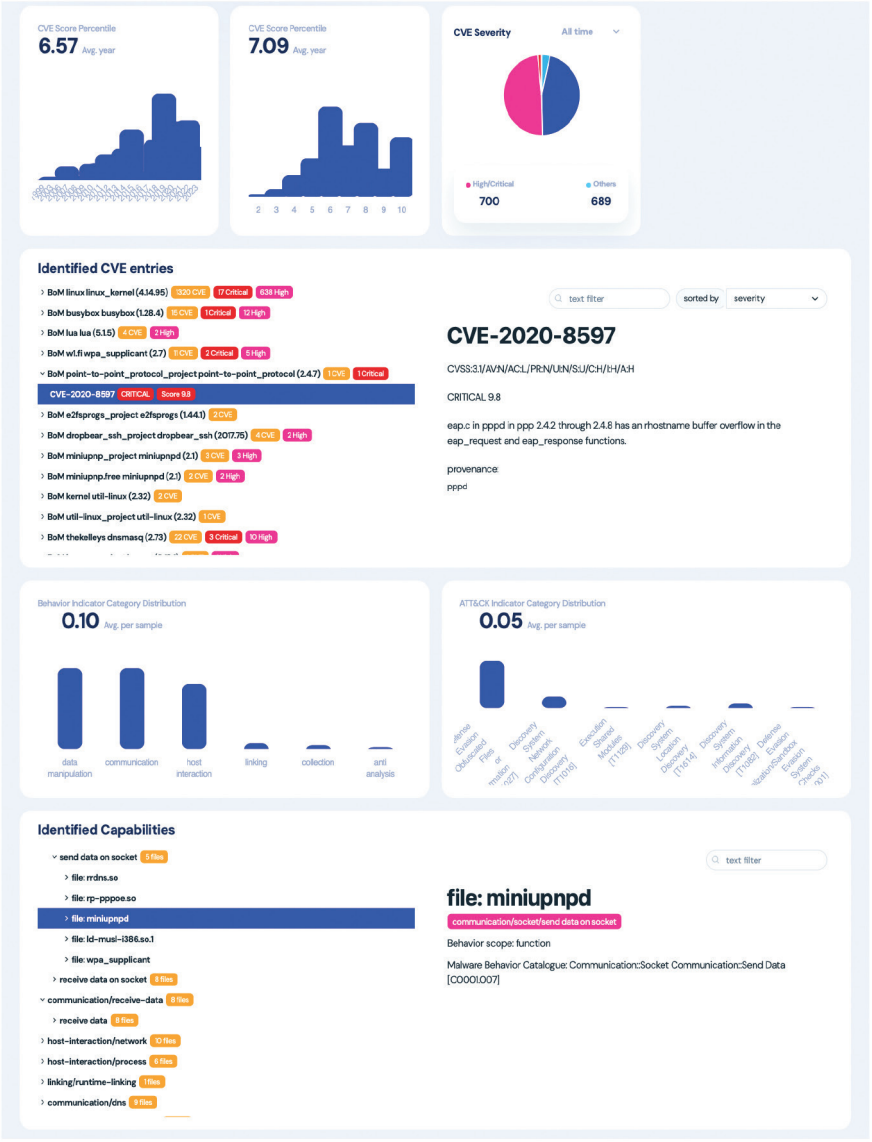
ERS0 organizes firmware images based on the concept of a repository. A repository, in this context, embodies a collection of firmware images earmarked for management and analysis. It is essential to note that each repository functions independently from the others. These repositories are exclusively owned by the current users and upcoming features will enable access sharing.

The repository dashboard, shown in Figure 6, provides a high-level view of the security analysis within a repository of firmware images. Key metrics here include the number of binaries extracted (21,835), instances of CVE vulnerabilities (4,697), and the prevalence of high-severity vulnerabilities (48% of the total). The dashboard highlights the average severity score of CVEs (6.94), an increase in average severity score (+2.45% compared to some baseline), and the criticality of vulnerabilities, with 129 identified as critical severity (3% of the total). It also offers historical context with a line graph showing the average age of CVEs spanning over five years, suggesting how quickly vulnerabilities are being identified and potentially addressed. An analysis of behavior categories indicates a focus on root-level actions, while the ATT&CK tactics charts categorize observed behaviors and their sources, which are critical for understanding attack patterns and planning defenses.

FIGURE 6: ERS0 REPOSITORY DASHBOARD

Given a firmware image in the repository, the user can open its corresponding latest SBOM analysis report, as illustrated in Figure 7. The SBOM and vulnerability report provide an in-depth analysis of the vulnerabilities present in the system's software components. The CVE score percentile graphs illustrate the distribution of vulnerability severities over time, with an average yearly score of 7.10, pointing to the presence of significant security risks. The CVE severity pie chart reflects this by showing that high/critical vulnerabilities constitute a substantial portion (50 out of 83 total instances). The report lists several specific CVE entries, such as a critical vulnerability in the "BoM point-to-point_protocol_project" and a high vulnerability in "BoM lua." These entries are categorized by severity and offer actionable intelligence for prioritizing patches and remediation efforts. The data suggests a need to continuously monitor and update software components to mitigate these identified vulnerabilities effectively.

FIGURE 7: ERS0 ANALYTIC REPORT



The capability report delves into the specific functions and potential risks associated with files within the system. It categorizes behaviors into areas such as data manipulation, communication, and host interaction, with data manipulation being the most prevalent, averaging 0.10 instances per sample. This is exemplified by 35

files having data manipulation/encoding capabilities, with 27 of these using XOR encoding—a common technique for obfuscating data to evade detection. The file named "signify" is specifically identified as employing this technique, aligning with the Malware Behavior Catalog's [20] Defense Evasion category and the ATT&CK framework's T1027 indicator for obfuscated files or information. Such insights are crucial for identifying files that may pose a security risk using sophisticated obfuscation or evasion methods.

6. CONCLUSION

ERS0 offers an effective enhancement to firmware security and asset management processes. By integrating AI and large-scale language models, it provides an improved method for constructing SBOMs that can scale to accommodate a wide variety of string patterns and library signatures. The preliminary results suggest that ERS0 may offer more thorough SBOM coverage and a more accurate identification of vulnerabilities, while also bringing a broader set of features to the table. Its performance, when compared with open-source alternatives, indicates that it has the potential to support military operations with a scalable and efficient tool for managing firmware assets and securing the supply chain. Moving forward, ERS0 could play a supportive role in the continuous improvement of cyber defenses, catering to the evolving needs of military technology and infrastructure.

REFERENCES

[1] L. J. Camp and V. Andalibi, "SBoM vulnerability assessment & corresponding requirements," (response to Notice and Request for Comments on Software Bill of Materials Elements and Considerations), National Telecommunications and Information Administration, 2021.

[2] S. Cho, E. Orye, G. Visky, and V. Prates, *Cybersecurity Considerations in Autonomous Ships*. Tallinn: CCDCOE, 2022.

[3] E. D. Wolff, K. M. Growley, M. O. Lerner, M. B. Welling, M. G. Gruden, and J. Canter, "Navigating the SolarWinds supply chain attack," *The Procurement Lawyer*, vol. 56, no. 2, 2021.

[4] Fraunhofer FKIE. "fkie-cad/FACT_core: Firmware analysis and comparison tool." GitHub.com. Accessed: Mar. 12, 2024. [Online]. Available: https://github.com/fkie-cad/FACT_core

[5] Intel. "intel/cve-bin-tool: The CVE binary tool." GitHub.com. Accessed: Mar. 12, 2024. [Online]. Available: https://github.com/intel/cve-bin-tool

[6] "MITRE ATT&CK®." MITRE. Accessed: Mar. 12, 2024. [Online]. Available: https://attack.mitre.org/

[7] E-M-B-A. "e-m-b-a/embark: EMBArk—The firmware security scanning environment." GitHub.com. Accessed: Mar. 12, 2024. [Online]. Available: https://github.com/e-m-b-a/embark

[8] "unblob—extract everything!" Unblob.org. Accessed: Mar. 12, 2024. [Online]. Available: https://unblob.org/

[9] National Security Agency. "Ghidra." Ghidra-SRE.org. Accessed: Mar. 12, 2024. [Online]. Available: https://ghidra-sre.org/

[10] "Hex Rays—State-of-the-art binary code analysis solutions." Hex-Rays.com. Accessed: Mar. 12, 2024. [Online]. Available: https://hex-rays.com/ida-pro/

[11] Mandiant. "Mandiant/capa: The FLARE team's open-source tool to identify capabilities in executable files." GitHub.com. Accessed: Mar. 12, 2024. [Online]. Available: https://github.com/mandiant/capa

[12] J. H. Clark, D. Garrette, I. Turc, and J. Wieting, “CANINE: Pre-training an efficient tokenization-free encoder for language representation,” *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 73–91, 2022.
[13] D. Chicco, “Siamese neural networks: An overview,” *Artificial Neural Networks*, vol. 2190, pp. 73–94, 2021.
[14] Z. Fu, S. H. H. Ding, F. Alaca, B. C. M. Fung, and P. Charland, “Pluvio: Assembly clone search for out-of-domain architectures and libraries through transfer learning and conditional variational information bottleneck,” 2023, *arXiv:2307.10631*.
[15] W. W. Peterson and D. T. Brown, “Cyclic codes for error detection,” *Proceedings of the IRE*, vol. 49, no. 1, pp. 228–235, 1961.
[16] VirusTotal. “YARA—The pattern matching Swiss knife for malware researchers.” GitHub.com. Accessed: Mar. 12, 2024. [Online]. Available: https://virustotal.github.io/yara/
[17] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
[18] Red Canary. “Atomic-red-team: Small and highly portable detection tests based on MITRE’s ATT&CK.” GitHub.com. Accessed: Mar. 12, 2024. [Online]. Available: https://github.com/redcanaryco/atomic-red-team
[19] T. Nowak. “NTAPI undocumented functions.” NTinterlnals.net. Accessed: Mar. 12, 2024. [Online]. Available: http://undocumented.ntinternals.net/
[20] MITRE. “Malware behavior catalog (version 3.0).” GitHub.com. Accessed: Apr. 8, 2024. [Online]. Available: https://github.com/MBCProject/mbc-markdown

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

Legal, Policy, and Compliance Issues in Using AI for Security: Using Taiwan's Cybersecurity Management Act and Penetration Testing as Examples

Wei-Che Wang
National Institute of Cyber Security
Taipei, Taiwan
wayne@nics.nat.gov.tw

Abstract: As artificial intelligence (AI) technology advances rapidly, integrating AI into cybersecurity practices poses new challenges for professionals. This paper focuses on the legal and policy implications of employing AI tools in penetration testing (PT). Key issues explored include liability in cases where AI tools cause damage and legal compliance challenges for organizations mandated to conduct PT. This paper argues that in the case of Taiwan, a comprehensive consideration of relevant laws, such as the Code of Civil Procedure, will be needed as AI products and tools become more widespread. The other issue concerns defining qualified PT, using Taiwan's Cybersecurity Management Act as an example. This paper concludes that, in addition to proper AI governance, governments should consider the legal frameworks necessary for the practical application of AI products or systems and develop appropriate AI safety testing methods to offer reference guidelines for public agencies to introduce risk-controllable AI tools, thus preparing for the transition into the AI era.

Keywords: *AI for cybersecurity, penetration testing, legal compliance, AI policy, product liability, Taiwan's Cybersecurity Management Act*

1. INTRODUCTION

As artificial intelligence (AI) technology advances, its relationship with cybersecurity has become increasingly noteworthy. Researchers have recently explored the potential of AI to automate cyber attacks. Some studies have focused on AI trained through machine learning methods, indicating that, while such AI might not currently revolutionize cyber attack techniques, it can effectively enhance the efficiency of each step in the Cyber Kill Chain.[1] Other researchers have used deep learning to investigate the performance and feasibility of AI tools in conducting automated penetration testing.[2] This suggests that AI can be not only a powerful tool to increase the threat of cyber attacks but also a potential means of bolstering an organization's cybersecurity defenses.

Market research shows that the value of AI cybersecurity products is still rising, a trend driven, not surprisingly, by the escalating severity of cyber attacks.[3] This increase in cyber threats, closely linked with the maturation of technologies like 5G cellular networks and IoT (the internet of things), compels government agencies, enterprises, and even individuals to allocate more resources for cybersecurity. For instance, large-scale data breaches can lead to significant financial losses and damage a company's reputation. Among various cybersecurity products, AI has emerged as a crucial technology in solutions, speeding up the identification of and response to cyber threats. Consequently, cybersecurity products augmented with AI technology are gaining popularity in the market.

However, throughout its development, AI has generated controversy. Issues raised include the lack of algorithmic transparency; vulnerability to cyber threats; potential discrimination in decision-making; contestability in AI decisions; the legal status of AI; intellectual property rights issues in AI; impact on labor, employment, and economic matters; privacy and data protection; accountability for damages caused; and lack of mechanisms for risk accountability.[4] These controversies have garnered significant attention and debate in the past. With the widespread adoption of neural network methodologies, also known as "black boxes," understanding how AI arrives at a specific answer or decision has become increasingly challenging. Today, as AI applications become more extensive and varied, the importance of these issues grows, necessitating more urgent attention and resolution.

1 Ben Buchanan et al., Automating Cyber Attacks (Center for Security and Emerging Technology 2020), https://doi.org/10.51593/2020CA002.

2 Zhenguo Hu et al., Automated Penetration Testing Using Deep Reinforcement Learning (2020), https://www.jaist.ac.jp/~razvan/publications/automated_penetration_testing_reinforcement_learning.pdf.

3 *Artificial Intelligence (AI) In Cybersecurity Market Size USD 102.78 BN by 2032*, NASDAQ OMX's News Release Distribution Channel (Jan. 23, 2023), https://www.proquest.com/wire-feeds/artificial-intelligence-ai-cybersecurity-market/docview/2768121329/se-2.

4 Rowena Rodrigues, Legal and Human Rights Issues of AI: Gaps, Challenges, and Vulnerabilities (2020), https://doi.org/10.1016/j.jrt.2020.100005.

As AI applications become more widespread, an increase in related legal disputes is anticipated. According to a study published by the Stanford Institute for Human-Centered Artificial Intelligence, in 2022, the United States saw 110 AI-related litigation cases, 6.5 times more than in 2016.[5] Of these, 29% were civil law cases, 19% were related to intellectual property rights, and 13.6% were contract law. Currently, civil cases greatly outnumber criminal or national security-related cases.[6] As AI increasingly impacts people's lives, the number of ensuing legal disputes will rise accordingly.

For instance, the penetration testing operations discussed in this article inherently carry certain security risks for the tested systems. If AI-driven testing leads to property damage or, more gravely, endangers human life, determining the legal relationships involved and allocating responsibility becomes a critical issue. This paper explores these aspects by examining AI policies and legal frameworks in major countries today.

Different governance methods can be chosen depending on the required purpose or degree of enforcement, typically including policies, regulations, and reference guidelines. Legislation for emerging technologies must consider various aspects, such as the law's purpose, the subjects and scope under its regulation, how the law will be implemented, and its societal impact. Therefore, countries often allow new technologies to function and develop in society for a period of time to ascertain their ramifications before legislating. In the meantime, countries usually outline their approach to these technological issues through policies (such as national investment in development or encouraging public–private collaboration) and practical reference guidelines that provide interpretations and practical examples. This approach mitigates the impact of emerging technologies while allowing them to create more possibilities for overall technological advancement, transitioning society and legal systems smoothly from a regulatory vacuum to an established framework.

The legal and policy issues surrounding AI are currently in this transitional phase. As AI capabilities and applications continue to mushroom, many countries and international organizations have extensively discussed how to impose more legal obligations on AI (and its developers, trainers, or users). While many similar conclusions have been drawn, such as the need for trustworthy AI and algorithmic transparency, the question of whether to incorporate these conclusions into legal regulations and the potential impacts of such legislation are still being considered. Currently, National governments' approaches to AI legal policies can be broadly categorized into two types: "active legislative regulation" and "guiding free development." The former, exemplified by the European Union's AI Act, directly regulates subjects, AI systems, requirements, and legal effects (penalties). By contrast, the latter approach, observed

5 Artificial Intelligence Index Report 2023 (Stanford Institute for Human-Centered Artificial Intelligence 2023) 291.

6 *Id*. at 294.

in countries like the United Kingdom and the United States, guides the development of new technologies in a government-friendly direction without imposing restrictions on industrial and technological growth through significant policy documents and technical reference guidelines.

Irrespective of the approach taken in policy regulation, autonomous tools developed through AI training have substantially impacted people's lives. Hence, this paper focuses on the context of "conducting penetration testing," discussing the legal issues that may arise when autonomous AI tools make it more convenient to carry out such testing. Furthermore, it examines compliance challenges that institutions, organizations, or enterprises might face when incorporating AI products or services, using the requirement for regular penetration testing under Taiwan's Cybersecurity Management Act as an example.

2. PENETRATION TESTING WITH AI

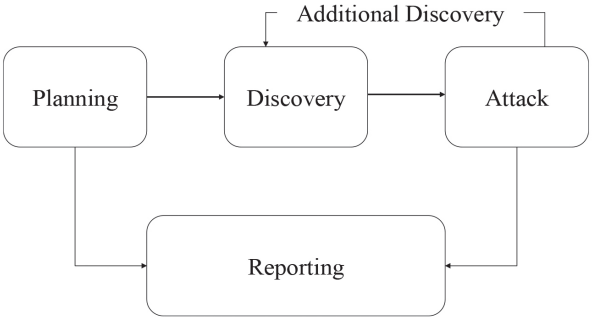
A. Introduction

Penetration testing is a method for assessing the security of information systems and detecting vulnerabilities. Tests are typically conducted by experienced cybersecurity experts. Testers simulate attacks using the same techniques and tools as attackers, often involving finding combinations of vulnerabilities on one or more systems. These combinations can grant more access privileges than would be possible through a single vulnerability, helping organizations (such as government agencies or businesses) improve their security defenses. By simulating attackers' behaviors, penetration testing can uncover vulnerabilities and weaknesses that many organizations are unaware of, thereby providing recommendations for improvement. Organizations can then use the tests' results and reports to better understand their system's security status, strengthen previously undiscovered weaknesses, and further protect their critical data and business operations.

Penetration testing can be divided into several key stages. This article explains these stages based on the guidance document of the National Institute of Standards and Technology (NIST) in the United States.[7] This document divides penetration testing into four stages: the Planning stage, the Discovery stage, the Attack stage, and the Reporting stage. The relationship between each stage is illustrated in Figure 1.

7 NIST, Technical Guide to Information Security Testing and Assessment (SP 800-115) (2008), https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-115.pdf.

FIGURE 1: STAGES OF PENETRATION TESTING



In penetration testing, the Planning stage sets the operational groundwork, including establishing rules, securing managerial consent, and defining clear objectives. This is crucial to safeguard testers, whose actions resemble legal hacking, and for organizations to differentiate between testing and actual cyber threats. The Discovery stage involves data collection and scanning, such as gathering the target's basic details (host names, IP addresses, system information, etc.), followed by vulnerability analysis, which uses this information to identify potential security gaps. Central to the testing, the Attack stage uses collected data to attempt system breaches, confirming vulnerabilities and their impact on system security. The approach adapts based on attack outcomes and additional information gathered, highlighting the interplay between the Discovery and Attack stages. The Reporting stage develops during other stages, culminating in a final report that outlines detected vulnerabilities and suggests reinforcement measures, building on the test plan's objectives and norms.

While penetration testing contributes significantly to enhancing system security, it can also potentially cause harm. For instance, simulated attacks might inflict actual damage on the system. Therefore, before conducting tests, testers should thoroughly understand the system architecture and develop contingency plans to reduce the likelihood of actual harm. Additionally, the testing process could expose sensitive information due to improper data or tool management or be exploited by hackers. Consequently, strict security measures must be implemented when the tests are carried out in order to ensure the tests' safety and confidentiality. Beyond technical controls, it is also essential to ensure that testers possess adequate professional skills and ethical integrity before testing. Testers should assess and manage potential risks to minimize their impact. In addition to controlling risks of actual harm, understanding the legal allocation of responsibility in the event of actual harm is crucial. This legal liability for any real harm should be managed through contractual arrangements or insurance.

B. The Theory and Practice of Autonomous Penetration Testing

Many tools now automate penetration testing through simulated environments by scanning and analyzing network structures and deployment environments and attempting attacks on known vulnerabilities. Manual penetration testing relies heavily on the tester's knowledge, experience, and skills, requiring significant time, effort, and resources. Autonomous AI tools, however, can streamline this process by autonomously exploring potential paths and weaknesses, analyzing intrusion strategies, and adapting to new information during the test. This automation can significantly lower the barrier to penetration testing and enhance efficiency.

Penetration testing includes various aspects such as software, hardware, environments, and personnel. While certain aspects, such as exploiting human errors or specific habits, may still rely on human experts, much of the testing, such as identifying vulnerabilities in software versions or system configurations, can be done autonomously. Key steps of penetration testing—target scanning, strategy formulation, and attack execution—can now be handled by autonomous tools shown in Table I, suggesting the feasibility of an integrated AI tool capable of conducting a complete penetration test with a single command. Thus, the concept of fully autonomous penetration testing is increasingly becoming a reality.

TABLE I: KEY STEPS AND CORRESPONDING TOOLS IN AUTOMATED PENETRATION TESTING

Steps	Description	Tools/methods used	Purpose/outcome
Reconnaissance	Scanning and detecting network topology	Nmap	Identify network structure and potential targets
Simulation	Simulating network architecture	CyberBattleSim	Understand the network environment and potential vulnerabilities
Strategic Planning	Planning penetration strategies	AutoPentest-DRL	Develop a strategic approach to penetration testing
Execution	Actual penetration testing operations	Metasploit	Carry out the attack to identify vulnerabilities
Reporting	Generating the test report	ChatGPT	Provide a detailed analysis and findings of the penetration test

With autonomous tools that can scan and detect the target network topology and the rapid advancement of AI technology, autonomous target penetration can be achieved. Even with current tool capabilities, autonomous penetration testing can integrate various AI tools for different tasks: Nmap for reconnaissance, CyberBattleSim for

network architecture simulation, AutoPentest-DRL for strategic penetration planning, Metasploit for actual operations, and large language models like ChatGPT for report generation. This concept of using a single AI tool for penetration testing is gradually becoming a reality.

The advantage of autonomous penetration testing tools is that they allow smaller organizations or businesses with limited resources to conduct thorough cybersecurity defense checks. This can significantly benefit national cybersecurity. However, the adoption of new technological tools should be approached with caution to avoid unforeseen risks.

C. Discussion of Legal Issues in Autonomous Penetration Testing

As previously explained, a consensus has yet to emerge on the regulatory framework for AI systems (or products, services, etc.). Given the global scale of the free market and the significance of major AI companies like Microsoft, Google, and Meta, it is crucial to closely monitor the regulatory developments in major economics like the European Union and the United States. These developments are likely to shape the direction of legal compliance for the entire AI industry. In response, Taiwan should closely observe international regulatory and policy trends, explore various emerging legal issues, and propose relevant legal and policy recommendations. This proactive approach aims to ensure a smoother integration of Taiwan's legal system with AI regulations and standards as AI becomes more widespread and its regulatory framework begins to take shape.

1) Exploring Liability for Damages Caused by AI Products

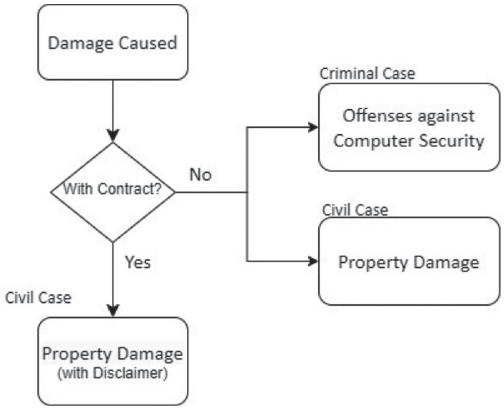
Over the past few years, AI has been extensively applied in numerous fields, such as autonomous vehicles and medical diagnostics. There has been considerable debate over whether AI can be considered a subject of tortious acts, that is, whether AI itself can be held responsible for damages caused. The prevailing opinion is that, under current legal principles, the answer remains unclear.

Accordingly, in the current legal system, until a new type of legal personhood for AI is defined, making AI systems responsible for the damages they cause, the primary subjects of liability are still those recognized as legal personalities under current law, namely, natural persons or legal entities. In the context discussed in this paper, corporations mainly involve the developers, providers, or suppliers of the AI tool, while natural persons are typically those using, operating, or carrying out the autonomous AI tools, issuing commands to carry out one or more stages of penetration testing.

When an actor uses the aforementioned autonomous tools and they result in damage to an organization, the relevant civil and criminal liabilities must be discussed. The

content involved in the planning stage of penetration testing plays a crucial role in making such a determination. That is, if written documents from the planning stage prove that the organization consented to the actor's hacking (testing) actions, it is easier to establish a contractual relationship to conduct penetration testing between the actor and the organization. This relationship, in addition to being used to determine whether the computer crimes under criminal law are "without cause," may also potentially exclude damages within a certain range from the compensation scope based on the agreement's substantive terms. The legal assessment is shown in Figure 2. In Figure 2, when damage is caused, the type of responsibility can be divided according to whether there was a contract. In a scenario that included a contract, the main responsibility would concern any damages that were caused outside the scope of the contract. In the absence of a contractual agreement, the victim can pursue both civil and criminal charges.

FIGURE 2: LEGAL ASSESSMENT OF THE DAMAGE TO INSTITUTIONAL INFORMATION SYSTEMS CAUSED BY AI TOOLS



In this scenario, the acting subject should be a natural person (for example, the engineer carrying out the testing project), and the autonomous AI program serves as a tool for conducting the test. The engineer is expected to supervise and intervene as necessary during the operation of the testing tool. For instance, actions such as immediate cessation, restoration, or repair should be taken if the autonomous tool successfully breaches a system, as this could lead to sensitive data leakage. Therefore, according to Taiwan's civil law provisions, tort liability is established if the autonomous AI tool causes property damage under the supervision and use of that natural person. Moreover, if it can be proven that the actor was negligent, they can be held responsible.

However, in most civil litigation cases, the allocation of the burden of proof substantially influences the outcome of the lawsuit. For example, according to Article 277 of Taiwan's Code of Civil Procedure, when a party asserts facts that are favorable to them, they have the responsibility to provide evidence for those facts. In other words, in the case scenarios discussed in this article, the party that suffered damage must prove that the AI tool caused the system's damage. Furthermore, to establish tort liability, they may also need to prove that the cybersecurity personnel responsible for supervising the use of the AI tool were negligent or worse.

Beyond civil liability, since penetration testing inherently involves acts of computer crime against information communication systems, relevant provisions can refer to Article 358 and subsequent articles of Taiwan's Criminal Code. Examples of such acts include "entering another's account and password, cracking computer protection measures, or exploiting computer system vulnerabilities to intrude into another's computer or its related equipment," "accessing sensitive information by obtaining, deleting, or altering the electromagnetic records of another's computer or its related equipment, causing damage to the public or others," or "using computer programs or other electromagnetic methods to interfere with another's computer or its related equipment, causing damage to the public or others." All these can constitute elements of computer crimes. At this point, whether the related intrusion actions have a legitimate reason for committing this "criminal act" is very important. This is also why the first penetration testing phase discussed in this article emphasizes the importance of project authorization documents.

2) EU's Product Liability Directives with AI

To develop trustworthy AI, the European Commission proposed a draft AI Liability Directive (AILD) in 2022,[8] which, along with the aforementioned AI Act, shapes the EU's legislative framework for AI. The primary purpose of the AI Liability Directive is to ensure that if users suffer harm due to AI products, the burden of proof for claims against AI is reduced. Additionally, clarifying how responsibility is allocated helps companies providing AI products or services to assess risks and reduce legal uncertainties. In line with the AI risk classification structure established by the AI Act, the new AILD applies in two scenarios. First, in claims for civil liability for negligence in non-contractual relationships, it requires disclosure of evidence concerning high-risk AI. Second, it adjusts the burden of proof in EU (member states') courts for compensation claims for damages caused by AI systems under non-contractual civil law.

In the section related to high-risk AI liability, this directive grants courts the power, under specific circumstances, to require relevant personnel of the high-risk AI (such as service providers) to disclose evidence related to the AI. To ensure fairness between

8 *Liability Rules for Artificial Intelligence*, European Commission, https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en.

both parties regarding evidence and information disclosure, the directive also sets many restrictions on the circumstances mentioned above. For example, the plaintiff must have already requested evidence from the AI-related personnel and been refused, and at the same time, the plaintiff must present sufficient facts and evidence to support the claim. On the other hand, the court should limit the scope of evidence disclosure to ensure that what is disclosed is relevant to the claim.

Although the regulations are more detailed in their applicability and conditions for high-risk AI, both the content of the regulations and the discussions during the legislative process indicate that the EU anticipates an increasing number of cases where AI products will conflict with human rights. Regardless of whether the current litigation system and the allocation of the burden of proof can properly handle such disputes, this AILD proposal may provide valuable references for Taiwan's legal and policy considerations.

At the same time, the EU is also considering amending the existing Product Liability Directive.[9] Since 2018, the European Commission has been amending the existing Product Liability Directive. The currently proposed revision has primarily updated three parts of the directive. First, it addresses the legally unclear concepts in the application of the law to emerging technologies. Second, it addresses the burden of proof that works against victims in cases involving products of emerging technologies, such as self-driving cars and AI products. Third, the previous Product Liability Directive had a threshold of €500 for claims, meaning that damages not reaching this amount could not be claimed; in the current proposal, this threshold has been removed.

The Product Liability Directive imposes responsibility on the economic operator of a product if a natural person suffers harm due to a defect in that product. In terms of enhancing the clarity of legal concepts, the new directive draft explicitly includes in the definition of "product" items commonly seen in the digital age, including digital files and software. Thus, AI-related products may also fall within this scope.

Regarding the requirements for the burden of proof, in addition to putting forward several presumptions where the causal relationship of damages is established, the new directive draft adds that if a plaintiff faces difficulty in proving the causal relationship between the product and the damage due to "technical or scientific complexity," the court may, under certain conditions (for example, if the plaintiff has provided sufficient evidence that the product is likely defective), acknowledge the causal relationship between the product and the damage for compensation. This legislative proposal echoes the approach taken in the AI Liability Directive, addressing the challenge of establishing tort liability after presenting strong evidence due to the high complexity

9 *New Liability Rules on Products and AI to Protect Consumer*, European Commission (Sep. 28, 2022), https://ec.europa.eu/commission/presscorner/detail/en/ip_22_5807.

of technology, which is difficult to handle with traditional legal concepts. Table II compares the two proposals.

TABLE II: COMPARISON OF THE EU'S AI LIABILITY LEGISLATION PROPOSALS

Aspect	AI Liability Directive	Product Liability Directive Amendment
Scope	High-risk AI and non-contractual civil liability claims	Includes products of emerging technologies, such as AI and self-driving cars
Legal Clarity	Addresses liability issues specific to AI, including complex AI systems	Enhances clarity of legal concepts for products in the digital age
Burden of Proof	Adjusts burden of proof in favor of victims in AI-related cases	Adds presumptions for the causal relationship in cases of technical/scientific complexity
Damage Threshold	None	Removes the €500 damage threshold for claims
Inclusion of Digital Products	Broader scope to AI-related products	Directly includes digital files and software

3) Legal Regulations Related to AI Liability in Taiwan

On the other hand, when it comes to Taiwan's liability-related regulations, discussions regarding property damage caused by the use of autonomous AI products can be approached from both substantive and procedural law perspectives. Taiwan's legal system is deeply influenced by the civil law tradition. The provisions of substantive law mainly involve civil law and consumer protection law. Additionally, when the damage pertains to the use or leakage of personal data, the Personal Data Protection Act might apply. On the procedural law side, because the burden of proof comes into play in litigation, it is necessary to consider whether civil litigation laws should be adapted to reduce the plaintiff's burden of proof in lawsuits concerning disputes over AI product liabilities.

In Taiwan's civil law, the issue of compensation for property damage needs to be addressed. When AI tools are used for penetration testing and this results in damage to the tested host system, given the frequent information asymmetry between product manufacturers and consumers in terms of economics, knowledge, and product performance, Taiwan has established the Consumer Protection Act[10] to safeguard consumer interests. This act specifically addresses business operators involved in designing, producing, and manufacturing goods or providing services. When these goods enter the market or services are provided, business operators must ensure the safety of these goods or services, according to the reasonably expected standards of

10 Consumer Protection Act, art. 7, https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=J0170001.

current technology and professional expertise at that time. The burden of proof for claims regarding these facts also rests with the business operators.

Under the current provisions of Taiwan's civil litigation law, a party who asserts facts favorable to their case bears the burden of proof (Article 277). However, recognizing that not all types of disputes can be adequately addressed by this general rule, the law provides a flexible adjustment mechanism with the provision: "This limitation does not apply if there are specific legal stipulations to the contrary, or if adhering to this rule would manifestly be unfair."

A classic example of a type of dispute in which the burden of proof is allocated differently is medical disputes. Other types include environmental pollution and traffic incidents; there, Taiwan's legislation also explicitly includes the responsibility of product manufacturers. In medical litigation, due to the high level of expertise, uncertainty, unpredictability, and information asymmetry inherent in medical practices, and given that the general public lacks relevant professional knowledge, it is often challenging to prove negligence in medical acts by hospitals or physicians, especially since medical records and equipment are predominantly controlled by them. Therefore, a shift in the burden of proof is applicable in these cases.

Damages caused by AI products share similar characteristics. In autonomous penetration testing, because AI models function like a "black box" (meaning that the reasons for their decisions cannot easily be discerned), operators, despite their own IT or cybersecurity expertise, may rely on the AI's judgment. Critical information about the AI model's training, judgment, or decision-making logic is usually held by the operators or suppliers who trained and adjusted the AI product.

D. Legal Implications of AI in Penetration Testing

The legislative model of the EU demonstrates that the governance policy for AI and the subsequent design of the responsibility structure must be considered simultaneously. That is, the categorization of AI systems must precede the question of whether to impose specific responsibilities on certain categories of AI.

As the above discussion shows, in terms of legal liability arising from damages caused by using AI tools for penetration testing, Taiwan already has applicable provisions for both civil and criminal liabilities. However, it may still be necessary to adjust the relevant regulations based on the characteristics of AI tool products, such as the issue of the burden of proof in civil damage compensation lawsuits. Additionally, in terms of administrative responsibility, we can refer to the legislative model of the EU AI Act concerning roles such as manufacturers or suppliers of AI tools, products, or services.

After establishing the basic legislation for Taiwan's Artificial Intelligence Act, we can then continue with special legislation to regulate these important obligations.

3. DISCUSSION OF LEGAL COMPLIANCE ISSUES OF AUTONOMOUS PENETRATION TESTING

A. The Cybersecurity Management Act and the Role of Penetration Testing

Given the practices in the cybersecurity industry and the provisions of Taiwan's Cybersecurity Management Act, it is not difficult to see that penetration testing is a relatively high-standard requirement in current cybersecurity defense testing. For example, the current Cybersecurity Management Act (and its related subsidiary laws, collectively called the CMA) requires organizations with a cybersecurity responsibility Level-C or above[11] to regularly conduct penetration testing on their core information communication systems. Additionally, for core information communication systems classified as "high" in protection level, the CMA also requires penetration testing during the development and acquisition stages of the Secure Software Development Life Cycle (SSDLC). Furthermore, according to the Enforcement Rules of the CMA,[12] agencies are required to include provisions for penetration testing when outsourcing their customized information and communication systems to ensure compliance and enhance security measures. This demonstrates the importance of penetration testing in compliance with Taiwan's cybersecurity-related legal requirements.

However, as this paper mentioned, a common issue agencies face in practice is that penetration testing requires considerable expenses and resources. If an agency decides to use autonomous tools for penetration testing and report generation, additional issues need to be addressed, such as whether these reports meet legal compliance requirements.

B. The Legal Effect of Penetration Testing

To answer the aforementioned question, it is necessary to explore what practical effects policymakers hope to achieve. Generally, if the process only involves simple scanning of a system for known vulnerabilities, it is usually referred to as a vulnerability scan. On the other hand, there may be exceptional cases in which professionals are hired for penetration testing but—due to negligence or other reasons—the results of their testing do not differ significantly from those obtained through mere tool scanning.

11 The cybersecurity responsibility levels of government agencies and specific non-government agencies are classified from high to low into Level-A, Level-B, Level-C, Level-D, and Level-E. Agencies rated Level-C and above are defined by regulations as those that maintain and operate, or outsource the establishment and development of, their cybersecurity systems. Regulations on Classification of Cyber Security Responsibility Levels, https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=A0030304

12 Enforcement Rules of Cyber Security Management Act, art. 4, https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=A0030303.

Therefore, this paper holds that, in defining what constitutes legally compliant penetration testing, one should observe whether the effects of the testing meet the needs of the agency's cybersecurity defense rather than merely determine whether tools were used for the test or whether there was human involvement.

As for what content meets an agency's cybersecurity protection needs, the "Solicitation Document for Government Agency Penetration Testing Service Outsourcing Proposal (Template),"[13] issued by the National Institute of Cyber Security in Taiwan, provides guidance. Besides listing all the necessary security testing items, the document makes two main points. The first is that the testers must have qualifications, such as cybersecurity certifications like CEH (Certified Ethical Hacker), CPENT (Certified Penetration Testing Professional), and so on. The second is that the submitted test report must be comprehensive. It should not only detail the methods used to discover vulnerabilities and the attack techniques employed and assess the risk level of the vulnerabilities but also—and this is most important for the agency—provide actionable improvement recommendations so that the agency can follow these to strengthen protection after the test.

Therefore, this paper holds that in the legal compliance issue of autonomous penetration testing, the judgment should be based on the effectiveness of the testing rather than merely on whether it is completed by autonomous tools. As for whether the use of autonomous tools (such as utilizing language models like ChatGPT to write vulnerability analysis reports) could lead to the leakage of the agency's sensitive information (for example, by inputting important core system configurations or internal network architecture information), that is another regulatory issue that needs consideration.

C. Balancing Automation and Risk

The impact of AI development is less about replacing humans than about assisting them—that is, making tasks that originally required many resources and had higher barriers easier to conduct or access. The penetration testing discussed in this paper is an example. If in the future, agencies start effectively using autonomous tools for penetration testing, that would be a positive development. However, this phenomenon requires attention from both the agencies themselves and the higher-level units conducting audits.

For the agencies themselves, although they might use AI tools for testing, they still need to be aware of the related risks, including damage to the information communication systems during testing or the leakage of sensitive information to AI tools, as mentioned earlier. To that end, this paper suggests that agencies should not

13 National Institute of Cyber Security, Solicitation Document for Government Agency Penetration Testing Service Outsourcing Proposal (Template), https://download.nics.nat.gov.tw/UploadFile/attachfilespmo/%E6%BB%B2%E9%80%8F%E6%B8%AC%E8%A9%A6%E6%9C%8D%E5%8B%99RFP%E7%AF%84%E6%9C%ACv5.0_1100915.pdf.

only thoroughly assess AI tools before selection and choose autonomous tools with lower risks but also hire operators with professional knowledge or qualifications.

Furthermore, auditors reviewing penetration testing reports provided by agencies will need to clearly understand the related background information of the report, including methods of execution, the tools used, test items, methods, scope, and improvement suggestions. That is, whether the testing complies with regulations is independent of whether it was conducted manually or automatically. Judgment should still be based on the substantive content of the testing.

4. CONCLUSION

The legislative model of the EU demonstrates that the governance policy for AI and the subsequent design of the responsibility structure must be considered simultaneously. That is, the categorization of AI systems must precede the question of whether to impose specific responsibilities on certain categories of AI. On the other hand, Taiwan's experience with cybersecurity legislation shows that the increasingly powerful performance of AI tools will also affect compliance with existing regulations or the auditing of standards.

In light of the associated cybersecurity risks that may accompany the use of AI automation tools, this paper posits that while governments contemplate AI governance, they must also be attentive to ancillary approaches. For instance, with respect to the product liability of AI, as discussed in this paper, clear legal norms are needed that delineate product responsibilities. Moreover, for legal and compliance issues related to AI automation tools in various types of legal operations, standards or auxiliary guidelines should be established based on practical scenarios to address the impact of AI. Below, this paper also offers three recommendations.

First of all, AI tools should be accepted as assistance. AI's role is predominantly to simplify and make accessible tasks that were previously resource-intensive and complex. An illustration of this can be seen in the penetration testing discussed in this paper. Should agencies begin to effectively deploy autonomous tools for penetration testing in the future, it would represent a significant advancement. However, such a shift demands vigilant oversight from both the agencies involved and the higher-level authorities responsible for auditing their activities.

Secondly, a list of usable AI tools should be established. While these tools are beneficial for testing, agencies must remain cognizant of potential risks, such as possible damage to information communication systems or unintended exposure of

sensitive data during the testing phase. Consequently, this paper recommends that agencies rigorously evaluate AI tools to select those with minimal risks and also ensure that they employ skilled operators who have the necessary expertise and credentials.

Finally, the validation of penetration test reports must prioritize the depth and quality of the content over superficial elements. It is imperative that auditors who review these reports from various agencies gain a comprehensive understanding of the detailed context provided within them. This includes not only the methodologies and tools employed but also the specific areas tested, the scope of the tests, and any recommendations for improvements. Importantly, the compliance of these tests with regulatory standards should be judged independently of the methods used, whether manual or automated. Decisions should be rooted in a thorough assessment of the actual findings and outcomes of the tests, emphasizing the importance of substance over form in these evaluations.

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

Not All Those Who Wander (Over the Horizon) Are Lost: The Applicability of Existing Paradigms of International Law to Cyberspace and the Interpretation of Customary International Law

Kristy Chan
LLM Candidate
University of Cambridge
BA Jurisprudence (Oxon)
kristychanwork@gmail.com

Joseph Khaw
BCL Candidate
University of Oxford
BA Jurisprudence (Oxon)
josephwykhaw@gmail.com

Abstract: It may be considered banal at this point for a State to assert that 'international law applies to cyberspace'. However, this belies tricky methodological questions regarding how a 'new' rule of customary international law (CIL) emerges. Cyberspace poses unique difficulties for the identification of CIL because of a paucity of publicly known State practice, vague statements, and attribution difficulties. However, this does not render CIL irrelevant to cyberspace. We argue that as the pace of technological development increases, interpretation of general rules of CIL may be used to ascertain their content when applied in cyberspace.

First, the proposed interpretive method is discussed. Second, State practice on the application of sovereignty and jurisdiction in cyberspace are considered to demonstrate interpretation in practice, focusing on extraterritorial botnet takedowns. Third, objections to the interpretive method are considered but shown to be ultimately unsustainable.

Normatively, the interpretation of CIL is an important tool for regulating cyberspace. First, it explains States' constant assertions that CIL applies to cyberspace despite the difficulties in meeting the usual tests. Second, on this approach, custom does not play catch-up to States' activities but develops contemporaneously. This allows

international law to peer over the horizon and be better prepared to tackle future challenges.

Keywords: *cyberspace, custom, identification, international law, interpretation, methodology*

1. INTRODUCTION

> *In their use of [information and communication technologies or ICTs], States must comply with international law … Hence, in current discussions, the question is no longer whether, but how international law applies to the use of ICTs by States.*[1]

> *Existing international law applies to cyber operations … Accordingly, the task of the International Groups of Experts … was to determine how such law applies in the cyber context.*[2]

This paper concerns methodology in customary international law (CIL). Specifically, if States think that international law applies in cyberspace, what does that mean from a methodological standpoint? *How* does a court work out what those rules are when applied? In what way, if at all, does that process differ from identifying a rule of CIL, which is 'to be looked for primarily in the actual practice and *opinio juris* of States'?[3]

We argue that applying existing international law to cyberspace can and should differ from identifying a new rule of CIL. Namely, it can be achieved through interpretation. Interpretation can help us look 'over the horizon' and enable custom to better address rapidly developing challenges in cyberspace, instead of playing catch-up. Ultimately, this article seeks to answer the call for Project 2100 through the domain of CIL,[4] strengthening custom as a tool for regulating cyberspace.

This paper proceeds in four sections. Section 2 explains our understanding of interpretation, its role in CIL methodology, and why cyberspace is a particularly

1 Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States, UNODA, A/76/136, August 2021, 17 (Brazil).

2 Michael Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (CUP 2017) 3 (*Tallinn Manual*).

3 *Continental Shelf (Libyan Arab Jamahiriya v Malta)* [1985] ICJ Rep 3 [27]; *North Sea Continental Shelf (Germany v Netherlands; Germany v Denmark)* [1969] ICJ Rep 3 [73].

4 See Sir Daniel Bethlehem KC, 'Project 2100 – Is the International Legal Order Fit for Purpose?' (*EJIL: Talk!*, 29 November 2022) <https://www.ejiltalk.org/project-2100-is-the-international-legal-order-fit-for-purpose/> accessed 6 January 2024.

ripe domain for applying the interpretive method. Section 3 applies interpretation to the rules of sovereignty in cyberspace, where State practice and *opinio juris* are sufficiently mature to deduce sub-obligations regarding extraterritorial botnet takedowns. Section 4 considers, but ultimately rejects, theoretical objections to interpretation. Section 5 concludes.

While this paper applies interpretation to sovereignty, the success of its broader argument regarding the potential of interpretation of CIL does not depend on accepting this example. Interpretation may be applied to various other aspects of international law in cyberspace. Nonetheless, space precludes comprehensive discussion of other areas.

2. WHAT IS INTERPRETATION?

A. Clearing the Field

Interpretation, as understood here, is a form of normative deduction in which 'new rules are inferred by deductive reasoning from existing rules and principles of CIL'.[5] This notion of 'interpretation' lies in the *application* of already recognized CIL to different factual matrices. While this process is guided by State practice and *opinio juris*, it is not the same as interpreting such practice itself. Understood this way, interpretation is irrelevant without an existing customary norm to be applied.

This might be confused with identifying the content of the existing norm, which critics of interpretation claim results in a vanishing line between identification and interpretation. Accordingly, interpretation allows courts to avoid the difficult requirement of establishing widespread and general State practice and *opinio juris*. Further, critics claim that customary rules do not exist in the abstract,[6] but are always tied to a particular context, such that 'identifying the content of the norm' and 'applying it to new contexts' are equivalent.

However, the relationship between the two is more like a Venn diagram. Interpretation and identification may overlap, and indeed in practice fleshing out the content of a rule as it applies in a new context looks the same as interpreting it to apply in a new context.[7] But in nascent areas such as cyberspace – where State practice and *opinio juris* are insufficient for the orthodox inductive methodology to bear fruitful results – interpretation and identification can produce different results. With knowledge of the

5 Stefan Talmon, 'Determining Customary International Law: The ICJ's Methodology between Induction, Deduction and Assertion' [2015] 26 EJIL 417, 423.

6 Massimo Lando, 'Identification as the Process to Determine the Content of Customary International Law' (2022) 42 OJLS 1042, 1049, 1051.

7 Dapo Akande, Antonio Coco, and Talita de Souza Dias, 'Drawing the Cyber Baseline: The Applicability of Existing International Law to the Governance of Information and Communication Technologies' (2022) 99 Intl L Stud 4, 18: 'In practice, there is little difference between [the] process of custom-identification and the interpretation and application of general customary rules to new phenomena.'

rationale behind the rules as well as some State practice and *opinio juris*, interpretation can be used in situations where one might describe existing practice as 'rather sparse, owing to the relative newness of the question', thus 'preclud[ing] the possibility of those conditions arising which are necessary for the *formation* of principles and rules of customary law'.[8] The proviso 'necessary' may suggest that we are off to a non-starter. However, its focus is on the *formation* of *new* customary rules, which do not concern us. Instead, we are concerned with interpreting *existing* customary rules when *applying* them to new circumstances.

For example, in Section 3, we acknowledge that one could argue that interpretation and identification both define the 'content' of the norm of sovereignty as it applies in cyberspace. However, interpretation allows a court to answer *how* the rule of sovereignty applies in cyberspace, whereas identification (in the absence of the requisite State practice and *opinio juris*) cannot. Critically, this can only be achieved by relying on an existing rule.[9] In this case, that is the rule prohibiting interference with another State's territorial sovereignty.[10]

B. Interpretation as Gap-Filler

In this approach, interpretation and identification play different roles and do not supplant each other. Talmon suggests that the court apply deductive – or at least non-inductive – methods of reasoning when faced with new contexts.[11] Cyberspace, we argue, is one such context. Indeed, while Akande, Coco, and Dias suggest that cyberspace is merely a bundle of information technologies and not a different 'domain' at all,[12] they underplay the difficulties that are unique to identifying CIL in cyberspace that make interpretation critical. First, there is currently insufficient State practice and *opinio juris*, and what does exist is too inconsistent. Second, even if more practice and *opinio juris* arise in the future, cyber-specific difficulties arise from (i) a paucity of publicly available State practice, given the secrecy surrounding national technology; (ii) vague statements making *opinio juris* unclear; and (iii) difficulties in attributing conduct.[13] On (ii), while *opinio juris* has always been elusive,[14] its elucidation in

8 *Delimitation of the Maritime Boundary in the Gulf of Maine Area* (Canada v US) (Merits) [1984] ICJ Rep 246 [81].

9 Talmon (n 5) 441: 'Deduction is the logically consistent extrapolation of the established body of CIL. It is, however, important that new rules of CIL are deduced only from existing legal rules or principles and not from postulated values.'

10 *Certain Activities Carried Out by Nicaragua in the Border Area (Costa Rica v Nicaragua) and Construction of a Road in Costa Rica along the San Juan River (Nicaragua v Costa Rica)* [2015] ICJ Rep 665 [93].

11 Talmon (n 5) 421–422.

12 Akande, Coco & Dias (n 7) 20.

13 Michael Schmitt and Stephen Watts, 'Collective Cyber Countermeasures' (2021) 12 HNSJ 373, 201–2; on attribution, see William Banks, 'Cyber Attribution and State Responsibility' (2021) 97 Intl L Stud 1039, 1046: 'Knowing the machines or IP addresses responsible for the hack is often difficult, costly, and time-consuming, and knowing those things does not necessarily lead easily to the responsible State.'

14 Omri Sender and Michael Wood, 'A Mystery No Longer? Opinio Juris and Other Theoretical Controversies Associated with Customary International Law' (2017) 50 Israel Law Review 299.

cyber contexts is even more difficult because of a lack of technical expertise from the actors to whom we usually turn to find *opinio juris*, such as State departments.[15]

Past instances of deductive reasoning by the International Court of Justice (ICJ) exhibit similar characteristics, as Talmon demonstrates.[16] This was the case where practice 'f[ell] short of proving the existence of a rule prescribing the use of equidistance, or any method, as obligatory',[17] as one might describe France's practice concerning botnet takedowns.[18] Similarly, in cases of negative practice consisting of omissions,[19] it may be 'practically impossible for one government to produce conclusive evidence of the motives which have prompted the action and policy of other governments'.[20] These considerations are all relevant in cyberspace.

One might question why a court should not just wait for further State practice and *opinio juris* to arise. First, as above, cyberspace is inherently inconducive to generating sufficient State practice and *opinio juris*. Thus, when a problem does come before the ICJ, unless the proposed methodological change is adopted, it may face a *non liquet*, which is 'no part of the Court's jurisprudence'.[21] Second, a critic might argue that there will be no *non liquet* if the closing rule in *Lotus* is applied: whatever is not prohibited is permitted.[22] However, as Hertogen has convincingly argued that *Lotus* does not stand for that proposition, the closing rule is of no help here.[23]

C. Interpretation as a Method of Legal Reasoning

But how can a court use interpretation to apply existing rules to new contexts? We propose the following elements of interpretation:[24] moving between levels of abstraction, teleological reasoning, and applying the rule.

15 Cf 'AI Safety Summit 2023' (GOV.UK) <https://www.gov.uk/government/topical-events/ai-safety-summit-2023> accessed 14 Apr 2024: The UK AI Safety Summit was intended to 'bring together international governments, leading AI companies, civil society groups and experts in research'. We agree that such events are valuable methods of elucidating more *opinio juris*, but we do not believe that it is sufficient given the quick pace at which novel technology in these fields develops.

16 *Gulf of Maine* (n 8) [81]; *Reparation for Injuries Suffered in the Service of the United Nations* [1949] ICJ Rep 174, 182 in Talmon (n 5) 422.

17 *Continental Shelf (Libya/Malta)* (n 3) [44].

18 Jack Kenny, 'France, Cyber Operations and Sovereignty: The "Purist" Approach to Sovereignty and Contradictory State Practice' (*Lawfare*, 12 March 2021) <https://www.lawfaremedia.org/article/france-cyber-operations-and-sovereignty-purist-approach-sovereignty-and-contradictory-state-practice> accessed 4 January 2024.

19 Paul C Ney Jr, 'DOD General Counsel Remarks at US Cyber Command Legal Conference' (*U.S. Department of Defense*, 2 March 2020) <https://www.defense.gov/News/Speeches/speech/article/2099378/dod-general-counsel-remarks-at-us-cyber-command-legal-conference/> accessed 7 January 2024: 'There is not sufficiently widespread and consistent State practice resulting from a sense of legal obligation to conclude that customary international law generally prohibits … non-consensual cyber operations in another State's territory.'

20 *North Sea Continental Shelf* (n 3) 246 (Dissenting Opinion of Judge Sørensen).

21 *Legality of the Threat or Use of Nuclear Weapons* (Advisory Opinion) (Dissenting Opinion of Judge Higgins) ICJ Rep 226 [36]; Talmon (n 5) 23.

22 *The SS 'Lotus'* 1927 PCIJ Series A, No 10, 18.

23 An Hertogen, 'Letting *Lotus* Bloom' [2016] 26 EJIL 901, 903.

24 Andreas Kulick, 'Interpreting the Customary Rules on State Responsibility – Text, No Text, Hypertext' in P Merkouris, P Pazartzis and LA Sicilianos (eds), *The Rules of Interpretation of Customary International Law* (CUP 2025, forthcoming) 9.

Critics such as Lando suggest that interpretation is not a viable method of legal reasoning as a matter of practicality because there are too many potential rationales.[25] However, as Talmon points out, the inductive method is 'just as subjective, unpredictable, and prone to law creation by the Court as the deductive method'.[26] Indeed, Tassinis convincingly shows that the orthodox method involves 'interpretation at every step of custom's life'.[27] Our focus, 'interpretation in application', is only one such step. Even Lando concedes that where State practice and *opinio juris* are lacking and courts have greater discretion, 'the case for the interpretability of custom, framed as a means to limit the exercise of discretion in determining the content of customary rules, might be more compelling'.[28]

Messiness is not inherently objectionable in international law. Kulick has pointed to the 'Eton messiness' of CIL as a defining feature of it,[29] while the International Law Commission (ILC) has described treaty interpretation as 'a single combined operation', whereby different means of interpretation are 'thrown into the crucible'.[30] This is what it means for interpretation to be a true *method*, as 'methods do not necessarily predetermine answers; they help explain how they are reached'.[31]

3. INTERPRETING RULES ON SOVEREIGNTY AND JURISDICTION TO APPLY TO CYBERSPACE

Consider an example: how do the customary prohibitions against infringing a State's territorial sovereignty and extraterritorial enforcement apply in the context of an extraterritorial botnet takedown?

For example, in the Anonymous Sudan botnet attack in November 2023, in which a Russia-backed group targeted networks in the US and Europe, any cross-border enforcement by States against Anonymous Sudan might be considered a breach of the targeted State's sovereignty. Thus, an attempt to 'delete' the webshells of a botnet attack – as States did during the takedown of EMOTET, which involved inserting malware into unknowing users' computers and initiating delete sequences[32] – would be considered an internationally wrongful act entailing State responsibility.

25 Lando (n 6) 1056.
26 Talmon (n 5) 432.
27 Orfeas Chasapis Tassinis, 'Customary International Law: Interpretation from Beginning to End' [2020] 31 EJIL 235.
28 Lando (n 6) 1046.
29 Kulick (n 24) 16; Section 4.
30 ILC, *Report on the Work of the Sixty-Eighth Session*, Subsequent Agreements and Subsequent Practice in Relation to the Interpretation of Treaties, Draft Conclusion 3(5) (UN Doc. A/71/10 (2016) 120).
31 Christian Tams, 'Self-Defence against Non-State Actors: Making Sense of the "Armed Attack" requirement' in Anne Peters and Christian Marxsen (eds), *Max Planck Trialogues on the Law of Peace and War* (CUP 2019) 93.
32 Daniel Rosenberg, 'Seizing the Means of Disruption: International Jurisdiction and Human Rights in the Expanding Frontier of Cyberspace' (2022) 55 NYU J Intl L & Pol 125, 143.

This is, however, at odds with current practice, where States hack into computers even where their location is unknown,[33] resulting in a potential breach of the victim State's sovereignty. States appear to have taken this uncertainty as a 'grant of jurisdiction', resulting in a 'new paradigm of enforcement jurisdiction'.[34] Further, not only are these operations announced *ex post facto*, kept quiet, or intentionally obfuscated such that *opinio juris* is difficult to find,[35] but State practice is also conflicting and disparate.[36] Even though many operations have highlighted their collaborative nature,[37] because the endpoint of the hack is not known until after the operation is conducted, acting States cannot claim to have obtained the victim State's consent.[38] This is exacerbated by the use of masking tools such as the dark web, which may obfuscate the true host of any botnets and hence render the 'endpoint' of the law enforcement agency's action unknown before conducting the cross-border operation.

Finally, existing treaties, such as the Budapest Convention, are of no help regarding botnets.[39] The Second Additional Protocol to the Convention on Enhanced Co-operation and Disclosure of Electronic Evidence (CETS No. 224) does provide for 'emergency mutual assistance',[40] but the existence of a treaty rule does not itself preclude a customary rule on the same matter, though the treaty may contribute to the backdrop against which a particular customary rule is interpreted.[41]

This is not merely of academic interest: botnets have caused billions of dollars in damage[42] and implicate other extraterritorial enforcement operations against, for example, child pornography rings. While these might, like botnet takedowns, be benign acts that international law can choose not to regulate,[43] the lack of an international legal

33 ibid 144–48.
34 ibid 132, 142.
35 ibid 144.
36 See Kenny (n 18).
37 Office of Public Affairs of the US Department of Justice, 'Qakbot Malware Disrupted in International Cyber Takedown' (*U.S. Department of Justice*, 29 August 2023) <https://www.justice.gov/opa/pr/qakbot-malware-disrupted-international-cyber-takedown> accessed 9 March 2024.
38 Rosenberg (n 32) 148.
39 ibid 132: The Budapest Convention does not satisfactorily address extraterritorial enforcement of cybercrime laws and was drafted before the cloud era, when data was stored primarily in States' servers and not overseas, which fails to recognize the 'sheer mass of data transmitted across borders'. See also the Council of Europe's Explanatory Report and Guidance Notes (2022) at 305 on art 32b, suggesting that certain situations of transborder access of data by law enforcement officials are 'neither authorized nor precluded'.
40 Council of Europe, 'Explanatory Report to the Second Additional Protocol to the Convention on Cybercrime on Enhanced Co-operation and Disclosure of Electronic Evidence' (2022).
41 Katie Johnston, 'The Nature and Context of Rules and the Identification of Customary International Law' (2021) 32 EJIL 1167.
42 Rosenberg (n 32) 127.
43 *Accordance with International Law of the Unilateral Declaration of Independence in Respect of Kosovo* (Advisory Opinion) [2010] ICJ Rep 403, 478 (Opinion of Judge Simma) [9]: international law might be 'deliberately neutral or silent' on a particular issue, so 'an act might be tolerated [but that] would not necessarily mean that it is legal, but rather that it is not illegal'.

framework may result in 'the cure be[ing] worse than the disease'.[44] For example, States may abuse this to pursue 'active cyber defence'.[45]

A. Changing Levels of Abstraction

As a form of deductive reasoning, interpretation requires moving from the general to the specific. However, Pomson argues that because the ICJ has refused to apply abstract precedents to more specific circumstances, interpretation is unworkable.[46] This is to be rejected. First, it is not true that the Court never deduces obligations by applying abstract precedents to specific situations.[47] For example, Talmon points out that in *Corfu Channel*, the UK argued for the existence of a peacetime obligation using State practice during wartime as precedent.[48] The ICJ generalized 'up' from the wartime precedent to the more abstract principles, and then 'down' to apply it to the minefield peacetime situation.[49]

Second, Pomson's examples are cases where States have chosen to plead based not on the 'applicability' of law but on the need for an exception to the existing rule. Thus, in *Jurisdictional Immunities*, the ICJ confined its analysis to 'acts committed on the territory of the forum state by the armed forces of a foreign state', rather than examining general precedents regarding torts committed on the forum State's territory.[50] However, the way Italy and Germany pleaded their case – as an exception rather than a limitation – meant that 'the Court was not free to adopt whatever analytical approach it saw fit, for example by framing the existence of the territorial tort exception as the interpretation of an existing customary standard'.[51]

The same argument applies to Pomson's example of *Arrest Warrant*. Indeed, he points out that because 'Belgium focused on whether an exception for war crimes and crimes against humanity existing regarding the immunity ratione personae … the Court … was essentially responding … on the very terms of that argument'.[52] In other words, there was no room for the Court to consider more abstract precedents or rules. In fact,

44 Rosenberg (n 32) 141.
45 Jack Goldsmith and Alex Loomis, *'Defend Forward' and Sovereignty*, Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 2102 (29 April 2021) <https://www.lawfareblog.com/defend-forward-and-sovereignty>.
46 Ori Pomson, 'Methodology of Identifying Customary International Law Applicable to Cyber Activities' (2023) 36 Leiden Journal Int'l Law 1023, 1041.
47 Talmon (n 5) 424: '[In *Corfu Channel*] the ICJ employed a triangular method of legal reasoning familiar in common law systems … where a precedent is similar to the case at bar in some important respects, but dissimilar in others, the [ICJ] identifies the general principles or rationale underlying the precedent and then decides whether this principle or rationale furnishes a suitable ground for deciding the case.'
48 ibid.
49 *Corfu Channel (UK v Albania)* [1949] ICJ Rep 4 [22].
50 *Jurisdictional Immunities (Germany v Italy)* [2012] ICJ Rep 99 [59].
51 Lando (n 6) 1052–53.
52 Pomson (n 46) 1037.

when it comes to State immunity, States have consistently framed their arguments in the form of exceptions to established rules.[53]

But if the perspective of States is what matters,[54] then this argument does not hold in cyberspace. States *do* view the content of norms from the top down and seek to establish how it applies in cyberspace, rather than seeking to establish a separate and parallel rule or an exception to the existing rule. The door is open to interpretation.

B. Teleological Reasoning

The next step is to determine whether the rationale for the main rule applies to the new situation. While Lando has argued that teleological reasoning is circular because the content of the existing general rule is itself the rationale, this is not true. This is because sovereignty and jurisdiction are residual rules of international law.

Residual rules are to be contrasted with norms such as humanitarian intervention, which operates as an exception to an existing rule. Here, CIL is not a 'micro-manager' but 'fills lacunae' when 'the diverse rules adopted by States collide'.[55] Thus, sovereignty is 'a residual rule that applies when no clear rule either prohibits or permits an action'.[56] It is up to the ICJ to decide this residual rule by considering the practice and the ultimate rationale of that area in question. In *Lotus*, the Court did not find that there was a 'presumption of freedom': it merely rejected a 'presumption against freedom', indicating that there may be limits on the exercise of sovereignty even when there is no express prohibition.[57] According to Hertogen, the rationale for this residual rule was that in the context of jurisdiction, 'territorial sovereignty must be exercised to ensure coexistence between independent States'.[58] The rule produced from this was that enforcement jurisdiction was prohibited unless a permissive exception could be established, such as consent.

It follows that when considering the applicability of sovereignty to cyberspace, the ICJ must continue to evaluate what is required for States to 'peacefully coexist', and it is *legitimate* to do so. According to *Lotus*, rules on enforcement jurisdiction are strictly controlled because of the horizontality of international law and the equality of States, regardless of size or history. Thus, the ICJ may start from the position that *any* infringement of territorial sovereignty entails a violation of sovereignty – that

53 Eg *Alleged Violations of State Immunities (Islamic Republic of Iran v Canada)* [2023] (ICJ proceedings instituted by Iran against Canada (27 June 2023), in which Canada argues for a 'terrorism' exception to State immunity).

54 Johnston (n 41) 1172: 'When the identification of [CIL] occurs in the context of litigation, much will therefore depend on how the issue is argued by the parties and how the rules involved are ultimately characterized by the Court … There does not appear to be any case before the ICJ where a party has succeeded in an argument relying on a customary rule that has been characterized as an exception to an existing customary rule.'

55 Hertogen (n 23) 911.

56 ibid 911.

57 ibid 908.

58 ibid 910.

is, identifying the existing rule. A court might choose to apply this *mutatis mutandis*, accepting that any intrusion into a State's cyberspace would result in a violation of its sovereignty. This would itself involve taking a stance on the rationale of the rules on sovereignty and jurisdiction – as the Permanent Court of International Justice did in *Lotus*. Alternatively, the ICJ may consider that the meaning of this changes given the porous nature of cyberspace and the need to address cybercrime in an increasingly interconnected world. In this approach, not *every* infringement of territory by cyber means entails violating the victim State's sovereignty.

C. Applying the Rule

However, the ICJ must choose between competing rationales and concretize the rule in application. This is the most controversial part of interpretation, especially compared to the inductive method, which assumes that one only needs to 'add up' State practice and *opinio juris*.[59]

For example, the ICJ may choose to adopt a 'de minimis' approach, where there is no violation of sovereignty if the effects of the State's hackback are minimal and the means used are the least intrusive. This is based on the need to ensure peaceful coexistence between States (the rationales outlined above), which, in a cyber context, necessitates some degree of jurisdictional overlap. However, Rosenberg has persuasively argued that this is only a good way of regulating an operation that has already taken place, and further restrictions are necessary to constrain State activities here.[60] Alternatively, the ICJ may consider that the absolute territorial prohibition is mirrored here and that any such hackback amounts to a violation of the hacked State's sovereignty. This may be motivated by considerations of undermining sovereignty in non-cyber domains.[61]

The point is that all of these interpretations are open to the ICJ – but it must choose, as it did in *Lotus* and *Nuclear Weapons*. In the latter, it evaluated existing practice and found that the general practice was prohibitory. Thus, proof of an exception – a permissive rule – had to be established, like in *Jurisdictional Immunities*. In cyberspace, the doors are wide open: it is more like *Lotus*, where Turkey and France disagreed on whether the content of sovereignty was permissive or prohibitive. It is

59 Moises Montiel, 'Fantastical Opinio Juris and How to Find It' (*Opinio Juris*, 23 June 2021) <https://opiniojuris.org/2021/06/23/fantastical-opinio-juris-and-how-to-find-it/> accessed 6 January 2024. Montiel argues that the two-element approach, where State practice and *opinio juris* is what CIL 'is', says nothing about how we get there; indeed, 'an equivalent would be saying that a cake is butter, flour, sugar, eggs, and milk. [This] is not wrong; but it adds nothing to the conceptual framework and hinders any attempt at identifying how to bake the coveted delicacy.'

60 Rosenberg (n 32) 153.

61 Consider, for example, the African Union Peace and Security Council's most recent statement rejecting a *de minimis* approach to sovereignty in cyberspace: Russell Buchan and Nicholas Tsagourias, 'The African Union's Statement on the Application of International Law to Cyberspace: An Assessment of the Principles of Territorial Sovereignty, Non-Intervention, and Non-Use of Force' (*EJIL! Talk*, 20 February 2024) <https://www.ejiltalk.org/the-african-unions-statement-on-the-application-of-international-law-to-cyberspace-an-assessment-of-the-principles-of-territorial-sovereignty-non-intervention-and-non-use-of-force/ > accessed 9 March 2024. As a bloc of 55 States, this should be considered strong State practice pointing away from a de minimis approach.

up to the ICJ to decide what approach to take, albeit based on a canvassing of practice in the area.

It should be noted that these terms ('effects', 'means', 'substantive') are taken from State practice and *opinio juris*, which, as cyberspace is such a technical domain, are used to *guide* the Court's application of the existing rule to formulate the new one, rather than being used in the usual inductive sense. Thus, our method of interpretation is still guided by State practice and *opinio juris*.[62]

4. IS THIS STILL CUSTOM?

It is admitted that this method of approaching custom, while not entirely *lex ferenda*, cannot be said to be *lex lata*. However, there are still normative benefits to adopting this method that outweigh potential objections.

A. Objection from Principle

Most vocal among these objections is that custom produced by interpretation is simply not custom at all. If custom is a 'practice' that has 'general acceptance as law', how can it be up to judges to specify 'what' that practice is if there is simply *no practice*? For example, Pomson argues that 'the proposition that customary rules are interpretable suggests … that one need not "always" have reference to state practice and *opinio juris* to determine the content of a customary rule'.[63] The criticism, then, is that interpretation impermissibly adds something to the mix that renders it 'not' custom.

However, first, there is practice given that we are advocating for interpretation to be used when States say that a general norm applies – interpretation is a way for the Court to specify *how*. Second, it is open to the international community to adopt a more fluid understanding of custom, such as that espoused by Hakimi,[64] where even an argument about what custom 'is' at a given moment on a particular topic counts as doing 'custom'. Interpretation fits well into this canon, though space constraints preclude further in-depth discussion.

62 *Barcelona Traction, Light and Power Company, Limited (Belgium v Spain)* [1970] ICJ Rep 3 (Separate Opinion of Judge Jessup) [60]: 'No survey of state practice can, strictly speaking, be comprehensive and the practice of a single State may vary from time to time … However, I am not seeking to marshal all the evidence necessary to establish a rule of [CIL]. Having indicated the underlying principles and the bases of the international law … I need only cite some examples to show that these conclusions are not unsupported by state practice and doctrine.'

63 Pomson (n 46) 1031–32.

64 Monica Hakimi, 'Making Sense of Customary International Law' (2020) 118 Michigan Law Review 1487; Jutta Brunnée, 'Customary International Law Symposium: Making Sense of Law as Practice (*Opinio Juris*, 7 July 2020) <https://opiniojuris.org/2020/07/07/customary-international-law-symposium-making-sense-of-law-as-practice-or-why-custom-doesnt-crystallize/> accessed 6 January 2024.

A related criticism is that interpretation allows what the law *ought* to be to determine what the law is. However, the Court is no stranger to attempting to ensure that CIL keeps pace with modern realities. For example, when arguing in favour of recognizing a right to self-defence against non-State actors,[65] Judge Kooijmans emphasized the need to make rules on the use of force suitable for modern dispute resolution – notwithstanding that this required departing from the mainstream interpretation of the past 40 years. In this approach, interpretation fills the gap between customary rules and the real-life scenario before the Court and allows CIL to develop contemporaneously with States' activities.

B. Objection from Practicality

Does interpretation make custom too uncertain? Especially as we advocate for interpretation to be used in nascent, developing areas of law, we acknowledge this potential uncertainty. Emerging custom would thus appear to reflect the words of US Supreme Court Judge Cardozo that 'the law that governs between [S]tates has at times … a twilight existence during which it is hardly distinguishable from morality or justice, till at length the imprimatur of a court attests its jural quality'.[66] However, the point is that interpretation avails itself when all that exists is the existing customary rule that States have said applies to cyberspace, which is arguably even more uncertain.

C. Objection from a Lack of Consent

The final potential objection is that this fails to respect the need for the consent of States, especially non-Western States. This cherry-picking of State practice and *opinio juris* to 'guide' interpretation renders interpretation nothing more than judicial legislation in disguise.[67]

Consent may indeed be lacking because States may object to whatever rule is produced from the interpretive process. However, because we advocate for interpretation to be used as a last resort when there is insufficient State practice, consent is only *potentially* lacking. There is room for States to object to such interpretations or for the persistent objector doctrine to apply.[68] Further, Talmon has persuasively argued that the deductive method is compatible with consent, given that deduction relies on the application of *existing* legal rules.[69] Nevertheless, from the perspective of Third

65 *Armed Activities on the Territory of the Congo (Democratic Republic of the Congo v Uganda)* [2005] ICJ Rep 168 [10]–[15].

66 *New Jersey v Delaware*, 291 US 361 (1934), in Rudolf Geiger, 'Customary International Law in the Jurisprudence of the International Court of Justice: A Critical Appraisal' in Fastenrath, Geiger, Khan, Paulus, von Schorlemer & Vedder (eds) *From Bilateralism to Community Interest: Essays in Honour of Bruno Simma* (OUP 2011) 673, 683.

67 Georg Schwarzenberger, 'The Inductive Approach to International Law' (1965) 60 Harvard Law Review at 126–27.

68 For example, the UK has always held that sovereignty cannot be breached as a standalone independent primary rule of international law.

69 Talmon (n 5) 441.

World approaches to international law (TWAIL), it is conceded that our methodology prioritizes *opinio juris* of States that have expressed a view on how international law applies to cyberspace, which will often be States of the Global North. However, this is reason for us to encourage further expression by States to generate more material for interpretation to work with,[70] rather than rejecting interpretation *per se*.[71]

The final charge – that interpretation is nothing more than judicial legislation – must be firmly rejected. First, it has always been true that many actors contribute to the articulation of substantive standards of conduct.[72] Thus, the legally binding status of international law may be *justified* by the consent of States to be so bound, but their content is not merely an expression of that will. Second, today's international legal order has 'radically transformed' as formal processes of international law-making have slowed.[73] While Pauwelyn focuses on the contribution of transnational corporations and nonprofits,[74] we suggest that *courts* can also be part of this change. As technology develops, we 'require more flexible norms … continuously corrected to take account of new developments'.[75] One way of doing so is to permit a more flexible CIL methodology that equips courts to play a greater part in law development.

5. CONCLUSION

Our argument is ambitious but limited. It is ambitious in that we suggest interpretation can help custom become fit for purpose in the 21st century. It is limited in that we propose great limits on it: the *type* of norm in question must be amenable to interpretation, requiring a nuanced understanding of differences between areas of international law. Given the difficulties posed by cyberspace to the development of CIL – including a lack of publicly available State practice, vagueness in national statements, and the significant technical expertise required to understand rapid technological developments – we argue that interpretation is necessary to look over the horizon.

70 In Episode 8 of the online podcast *Jus Cogens*, Eric Jensen, one of the original drafters of the *Tallinn Manual*, emphasizes that the goal was to put forward what the drafters believed *was* the law (lex lata) that would be material for States to respond to. See Dan Efrony and Yuval Shany, 'A Rule Book on the Shelf? Tallinn Manual 2.0 on Cyberoperations and Subsequent State Practice' (2018) 112 AM J INT'L L 583, 588: 'The combination of silence and ambiguity in state practice and their reluctance to articulate their official policy in cyberspace prevents or, at least, slows the development of global norms of conduct.'

71 Jeffrey Kovar, 'The US' Practical Approach to Identifying Customary Law of Armed Conflict' (*EJIL: Talk!*, 21 August 2023) <https://www.ejiltalk.org/the-united-states-practical-approach-to-identifying-customary-law-of-armed-conflict/> accessed 6 January 2024.

72 ILC Draft Conclusions on the Identification of CIL (2018), Conclusion 4(3): '[The] conduct of other actors is not practice that contributes to the formation, or expression, of rules of customary international law, but may be relevant when assessing opinio juris.'

73 Joost Pauwelyn, Ramses Wessel and Jan Wouters, 'When Structures Become Shackles: Stagnation and Dynamics in International Lawmaking' (2014) 25 EJIL 733, 734.

74 ibid 741.

75 ibid 742–43.

Specifically, we argue that interpretation involves three stages: first, changing levels of abstraction; second, teleological reasoning; and third, applying the rule. Custom is not inherently opposed to any of these three stages. It is open for international law to choose interpretation as a methodology of CIL. However, on this approach, CIL does appear to be more interdisciplinary, less State-centric, and more contemporaneous. This is not to be rejected for fear of change. Indeed, 'the conceptual boundaries of how international law may look in the future are wide open'.[76] That surely includes methodological change.

76 ibid 734.

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

The Scope of an Autonomous Attack

Jonathan Kwik
Postdoctoral Researcher
T.M.C. Asser Institute
The Hague, Netherlands
j.kwik@asser.nl
j.h.c.kwik@gmail.com

Abstract: 'Attack' is an important term of art in international humanitarian law that serves as the basic unit of reference for many targeting obligations. It is often also asserted that human commanders of autonomous weapon systems (AWS) must make legal determinations 'per individual attack'. Divergent interpretations on what constitutes an attack nevertheless lead to drastically different conclusions with regard to the technology's lawfulness: interpreted narrowly ('each shot'), it precludes AWS technology entirely, while interpreted broadly ('each activation'), it sanctions extensive autonomous activity. This paper theorizes that *imprecision on the scope of attack* is an underappreciated aspect of the AWS controversy that hampers theoretical and diplomatic advancements. The legal boundaries of autonomous attacks are analysed through the lens of targeting law, and a scaling methodology is proposed that allows commanders to determine the maximum extent to which autonomous activity may still lawfully be grouped into one single attack. The paper argues that both overly narrow and broad interpretations are inconsistent with targeting principles and practice, instead favouring a middle-ground approach based on temporal and spatial proximity that properly respects international humanitarian law's (IHL) balancing philosophy between humanitarian and military interests. Through consideration of practical scenarios, the paper subsequently demonstrates how this impacts the application of targeting rules, such as at what intervals the commander's duty to verify or cancel is triggered and under what circumstances successive autonomous engagements may be grouped together for proportionality assessments.

Keywords: *attack, targeting, autonomous weapons, precautions, IHL proportionality, proximity in time and space*

1. INTRODUCTION

'Attack', defined in Additional Protocol I (AP I) Article 49(1) as 'acts of violence against the adversary, whether in offence or in defence', is an important term of art in international humanitarian law (IHL) to which many protections attach.[1] New technologies sometimes necessitate revisiting how 'attack' is interpreted. In the case of cyberweapons, their intangible nature provoked a shift from the traditional (physical) conception of *violence* to a more effects-based approach.[2] In contrast, the term has received less academic attention in the debate surrounding autonomous weapon systems (AWS). While there has been extensive discussion on whether AWS can be used to lawfully conduct attacks or can be designed to properly implement precautions,[3] it is usually presumed that the 'attack' component is relatively uncontroversial. The infliction of violence is an attack's *sine qua non*,[4] and as AWS are usually conceived as physical systems intended to inflict (physical) harm to military objectives,[5] there is little reason to doubt that employing an AWS constitutes an attack. It is also uncontended that commanders employing AWS are obligated to ensure that fundamental principles such as distinction, precautions and proportionality are upheld.[6]

However, there is a different and underappreciated point of legal uncertainty regarding the notion of attack as it relates to AWS – one that significantly impacts how targeting rules are applied to AWS attacks.

Consider the following scenario:

> **Scenario 1.** At 1200, Commander-A activated an AWS to attack a tank platoon, during which the system released seven shots at four tanks (successive shots were released because the AWS detected that the objective

1 MN Schmitt, '"Attack" as a Term of Art in International Law: The Cyber Operations Context' in C Czosseck, R Ottis and K Ziolkowski (eds), *4th International Conference on Cyber Conflict* (NATO CCDCOE 2012) 284.

2 See C Droege, 'Get off My Cloud: Cyber Warfare, International Humanitarian Law, and the Protection of Civilians' (2012) 94 International Review of the Red Cross 533, 552–557.

3 See eg WH Boothby, 'Highly Automated and Autonomous Technologies' in WH Boothby (ed), *New Technologies and the Law in War and Peace* (Cambridge University Press 2018).

4 Schmitt (n 1) 290.

5 Eg N Davison, 'A Legal Perspective: Autonomous Weapon Systems under International Humanitarian Law', *UNODA Occasional Papers No. 30* (2017) 5. There is no consensus on how to define AWS. The term is used in this paper to refer to physical weapon systems enabled by artificial intelligence that can execute target selection and engagement independently. Whether this process actually occurs without human oversight is an operational choice made by the deploying commander.

6 There is universal consensus that targeting rules remain applicable to the use of AWS. See GGE on LAWS, 'Report of the 2023 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems' (24 May 2023) CCW/GGE.1/2023/2 para 21(a); M Pacholska, 'Autonomous Weapons' in Bartosz Brożek, Olia Kanevskaia and Przemysław Pałka (eds), *Research Handbook on Law and Technology* (Edward Elgar Publishing 2023).

was still operational after the initial strike). At 1500, the system was sent to attack four tanks guarding the city's four arteries, during which the AWS released six shots.

How many attacks has Commander-A launched today? That depends on the unit of measurement for 'attack':

- Each activation 2 attacks
- Each tank 8 attacks
- Each shot 13 attacks

This question is fundamental as 'attack' serves as the basic unit of reference for many targeting obligations.[7] 'Attack' is also used as a yardstick in many policy and diplomatic proposals. Take the proposition that a 'human should be in control of the system for each individual attack'.[8] This statement is actually not particularly controversial: precautionary obligations are addressed at 'those who plan or decide upon an attack',[9] and there is relatively broad international agreement that a machine cannot discharge legal obligations 'for' its human user (i.e., the commander).[10] However, depending on how broadly or narrowly one defines 'attack', this same proposition implies entirely divergent requirements for human involvement. At one extreme, we find interpretations that construe 'attack' at the narrowest level, such as at 'the stage when the munition is fired'.[11] This essentially 'precludes autonomous systems' altogether.[12] At the other end of the spectrum, we find positions arguing that an attack could potentially encompass the entire period in which the system is active (during which it may release multiple shots at many different objectives).[13] In concept, this permits autonomous operation at the 'trigger-pulling' level.[14]

The author of this paper perceives much of the disagreement in literature and debate regarding the lawfulness of autonomous technologies as derived from *imprecision and diverging interpretations on the scope of attack*. The aims of this paper are to explore how the scope of an AWS attack should be conceptualized legally based

7 HM Roff and R Moyes, 'Meaningful Human Control, Artificial Intelligence and Autonomous Weapons' (2016) Briefing Paper for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, 5.

8 T Chengeta, 'Defining the Emerging Notion of "Meaningful Human Control" in Autonomous Weapon Systems' (2016) 49 International Law and Politics 833, 875.

9 Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3 [AP I] art 57(2)(a).

10 GGE (n 6) para 21(c); ICRC, 'Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?' (2018) CCW/GGE1/2018/WP, para 32; US Department of Defense, *Law of War Manual, Updated July 2023* (US Department of Defense 2015) art 6.5.9.3.

11 W Boothby, 'Control in Weapons Law' in Rogier Bartels and others (eds), *Military Operations and the Notion of Control Under International Law* (TMC Asser Press 2021) 388.

12 ET Jensen, 'Autonomy and Precautions in the Law of Armed Conflict' in Rain Liivoja and Ann Väljataga (eds), *Autonomous Cyber Capabilities under International Law* (NATO CCDCOE 2021) 191.

13 Eg M Ekelhof, 'Moving Beyond Semantics on Autonomous Weapons: Meaningful Human Control in Operation' (2019) 10 Global Policy 343.

14 That is, each time a munition is released. See Jensen (n 12) 192.

on currently applicable targeting law and to propose a reasoned methodology to determine the appropriate scope of specific operational scenarios. Greater clarity on this legal question is beneficial both theoretically and practically. In terms of theory, it advances the doctrinal debate by offering greater precision as to the operational circumstances within which AWS would conflict with targeting requirements, such as proportionality and precautions. In terms of practice, the methodology provides field commanders with a readily usable cognitive framework with which to consistently determine how broadly or narrowly they may define 'attack' when AWS are planned for use during operations.

With this background established, the paper proceeds as follows. First, Section 2 elucidates the notion of *scope* as it relates to attacks under IHL and proposes a qualitative scale that allows us to theorize on the possible ways *Attack Scope* can be conceived. With this as the starting point, Section 3 considers the contours of how broadly or narrowly IHL permits attacks to be construed. It is argued that both overly narrow and overly broad conceptions of attack run counter to the purposes of IHL. Instead, a middle ground based on temporal and spatial criteria is recommended, which properly balances the humanitarian goals of IHL with the practicality of targeting. Section 4 applies this theory to targeting scenarios to demonstrate how commanders can use the proposed methodology to assess the appropriate scope of AWS attacks and how this impacts obligations in *attack* (in particular targeting rules in AP I Arts. 51 and 57). It also discusses the additional legal insights that this theorization provides. Section 5 concludes with overall remarks and recommendations.

Note that 'the commander, not the weapon system, makes legal determinations'.[15] This paper assumes that legal or moral agency cannot be assigned to machines. They can *perform functions consistent with* requirements such as distinction and proportionality,[16] but they cannot discharge IHL 'for' the human commander.[17] As such, the commander remains the primary person responsible for implementing precautionary obligations, and the steps proposed below are intended to be applied by a human commander prior to launching an AWS attack. This paper focuses on how the *user* (the commander) can ensure that obligations in *attack* remain respected when AWS are involved.

15 RJ Sleesman and TC Huntley, 'Lethal Autonomous Weapon Systems: An Overview' (2019) 1 Army Lawyer 32, 34. In making this determination, the commander may consult with a legal adviser, who helps them make a reasoned decision based on applicable law, the operational circumstances, and a risk assessment. Ultimately, however, the legal determination and responsibility remain the commander's. TD Gill and D Fleck (eds), *The Handbook of the International Law of Military Operations* (Oxford University Press 2010) para 31.02.

16 As such, no position is taken with respect to design requirements sometimes raised in literature, such as whether an AWS must be able to autonomously calculate proportionality or cancel attacks.

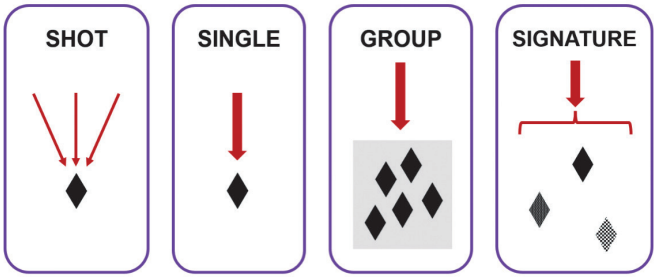
17 Above n 10.

2. DIFFERENT ATTACK SCOPES

First, let us clarify what is meant by *Attack Scope*. An attack has been variably described as 'combat action',[18] 'the commission of acts of violence',[19] or 'any military act of a violent nature'.[20] As indicated by the last quote, the threshold is very low: even a single sniper shot or lone dropped bomb qualifies as an attack.[21] There is also a clear limit at the other end of the spectrum: in any event, an attack is narrower than an operation.[22] Between these two extremes, however, there is less clarity on how one should determine the Attack Scope of a particular use-of-force instance, as Scenario 1 showed.

Figure 1 provides a visualization of the possible ways the scope of an AWS attack could be characterized.

FIGURE 1: ATTACK SCOPE SPECTRUM



Shot is the narrowest way Attack Scope can be construed. Here, each 'trigger-pulling action' by the AWS (i.e., each bullet shot or munition released) is considered a separate attack, even if these are successively aimed at the same objective. In contrast, *Single* considers all shots taken against the same target entity[23] as one attack.

Another possibility is to allow strikes against multiple entities to be combined. Under *Group*, all strikes against a Specific Target Group may be combined for legal purposes. The US Department of Defense defines a Specific Target Group as a 'discrete group of potential targets, such as a particular flight of enemy aircraft, a particular formation

18 Y Sandoz, C Swinarski and B Zimmerman, *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (Martinus Nijhoff 1987) ['AP Commentary'] para 1880.

19 *Prosecutor v Kunarac, Kovać and Vuković* (ICTY, Trial Judgment) IT-96-23-T & IT-96-23/1-T (22 February 2001) para 415.

20 Program on HPCR at Harvard University, *HPCR Manual on International Law Applicable to Air and Missile Warfare* (Cambridge University Press 2013) ['HPCR Manual'] 28.

21 M Bothe, KJ Partsch and WA Solf (eds), *New Rules for Victims of Armed Conflict: Commentary on the Two 1977 Protocols Additional to the Geneva Conventions of 1949* (Martinus Nijhoff 2013) ['2013 Commentary'] 329.

22 D Fleck (ed), *The Handbook of International Humanitarian Law* (Oxford University Press 2013) para 442.

23 'Entity' can refer to either a person or an object.

of enemy tanks, or a particular flotilla of enemy vessels'.[24] Therefore, to qualify as a Specific Target Group, the entities must share some proximity in both time and space. In Scenario 1, the platoon of four tanks targeted at 1200 hours can be considered a Specific Target Group. If characterized as a Group-level attack, all seven munitions released by Commander-A's AWS can thus be considered as a single attack.

The broadest approach is to consider all engagements against objectives with a shared *Signature*[25] to be part of a single attack. For example, a commander may activate an AWS to attack 'a particular model of tank or aircraft' within a specific area.[26] Compared to Group, the difference lies in the fact that these entities are not part of a *discrete* group as mentioned above. Many current in-use systems fall into this tier. 'Defensive' systems such as C-RAM, ground-based sentries and air defence use signatures such as trajectory, form and size to determine whether entities fall within the designated threat profile.[27] The Tomahawk, in so far it was launched over the horizon against a presumed presence of Soviet ships (but not any particular one),[28] would also qualify. As Scharre remarks with regard to the Harpy, its user 'does not know … *which particular* radars are to be engaged, only that radars that meet the Harpy's programmed parameters will be'.[29]

Using this scaling system allows us to analyse any use-of-force situation and establish the boundaries of where each attack legally ends and a new attack begins. To return to Scenario 1, *Shot* would consider that the AWS is engaging in a separate attack each time it detects an operational tank and releases a munition, even in iterative situations against tanks it previously engaged and failed to disable. Given that (human) commanders are required to engage in legal analysis prior to each attack (on the military nature of the objective and the proportionality of the engagement),[30] and granting that IHL 'precludes AWS from moving from one "attack" to another … without each individual

24 US Department of Defense, 'Autonomy in Weapon Systems' (2023) DoD Directive 3000.09 ['DoD Autonomy'] 23.

25 A signature is a 'pattern of sensor data that is taken to represent a target'. R Moyes, 'Target Profiles: An Initial Consideration of "Target Profiles" as a Basis for Rule-Making in the Context of Discussions on Autonomy in Weapons Systems' (Article 36 2019) 4. Landmines use weight to "decide" whether to explode. Modern AI systems can use complex signature combinations (eg heat, smoke, contextual semantic information) to determine whether an entity falls within the intended target set. See eg F Meng and others, 'Visual-Simulation Region Proposal and Generative Adversarial Network Based Ground Military Target Recognition' (2022) 18 Defence Technology 2083.

26 DoD Autonomy (n 24) 23.

27 M Ekelhof, 'Lifting the Fog of Targeting: "Autonomous Weapons" and Human Control through the Lens of Military Targeting' (2018) 71 Naval War College Review 61, 74.

28 See J Markoff, 'Fearing Bombs That Can Pick Whom to Kill' *New York Times* (11 November 2014) <www.nytimes.com/2014/11/12/science/weapons-directed-by-robots-not-humans-raise-ethical-questions.html> accessed 2 August 2023.

29 PD Scharre, 'Autonomy, "Killer Robots," and Human Control in the Use of Force' (*Just Security*, 9 July 2014) <https://www.justsecurity.org/12708/autonomy-killer-robots-human-control-force-part/> accessed 10 June 2021 (emphasis original).

30 A Cohen and D Zlotogorski, *Proportionality in International Humanitarian Law* (Oxford University Press 2021) 65.

attack being subject to human legal judgments',[31] this imposes a strict obligation on commanders to continuously monitor their AWS (and effectively precludes the lawful use of autonomous systems altogether).[32]

Single is less demanding. Consider the attack on the tank guarding the south artery at 1500 hours. Under this interpretation, the commander is required to confirm the military nature of the tank and proportionality *prior* to commencing their AWS attack (i.e., activating the system)[33] but is not required to repeat this assessment every time the AWS detects that another strike is necessary to disable the tank: they can leave the AWS to 'finish the job' independently and still be confident that they properly discharged all duties in attack. Compared to Shot, Single introduces slightly more risk with respect to the protections offered by IHL. For instance, what if, between shots, the tank unexpectedly rolls next to a market stall in an attempt to evade the AWS?[34] Improperly delineated, Single can also be taken to an absurd extreme. Suppose the commander orders an insurgent leader to be targeted based on biometric signatures, but the leader manages to dodge an initial attempt by the loitering system. Six hours later, he resurfaces again and is identified and killed by the AWS. Is this still part of the same attack? If not, at what point was the commander required to conduct their legal evaluations again?

Group and *Signature* are even more militarily efficient. For the 1200 sortie, construing the tanks as one group allows the commander to perform validation and the proportionality assessment only once for the platoon as a whole before activating the AWS. With regard to proportionality, imagine that for the 1500 sortie, the commander learns that attacking the north tank will unavoidably hit a market stall. Combining all four tanks for the purposes of military advantage might allow the commander to proceed with the attack on the north tank, compared to if this were classified as four Single-attacks.[35] Conversely, these broader conceptions of Attack Scope further amplify the risk to the civilian population, both by increasing the epistemic distance between the commander and the effects of the attack, and by potentially justifying significant magnitudes of collateral damage under the pretence that all incidental harm was inflicted 'under the same attack'.[36] The question thus arises: are there limits to how broadly or narrowly Attack Scope can be construed?

31 Roff and Moyes (n 7) 5.
32 Jensen (n 12) 191.
33 AP I (n 9) art 57(2)(a)(i) and (iii) respectively.
34 A fighter pilot in a similar situation would likely delay their re-strike for fear of excessive collateral damage.
35 See Section 4.
36 See Section 3.

3. DELINEATING ATTACK SCOPE

Moving between Attack Scope tiers involves trade-offs in protection versus practicality. The lowest tier (Shot) offers the greatest level of safety to the civilian population but is very demanding from a military perspective. It not only requires constant human supervision and legal evaluation per shot[37] but also construes the proportionality rule very narrowly (this limitation also applies to Single). Moving up the scale to Group and Signature, we see the opposite effect as efficiency becomes the dominant factor. Group allows commanders the benefit of only performing validation and proportionality evaluations once for a collective of objectives. Signature is yet more permissive since the objectives need not even constitute a disparate group. We can thus hold that from a military perspective, construing Attack Scope broadly is preferred. Those prioritising military efficiency will want to define the scope of their AWS attack as broadly as possible, as this confers practical and logistical benefits. Those emphasizing humanitarian interests will want to take the opposite position, and push for AWS attacks to be construed as narrowly as possible to maximize civilian protection.

Given these competing interests and the doctrinal ambiguity on this matter, there is clear value in establishing clear and reasoned guidance as to how far commanders may stretch the notion of *attack* when planning use-of-force using an AWS. To this end, this section first explores the theoretical contours that limit how narrowly or broadly Attack Scope *can* be conceived. Then, having identified these contours, more concrete criteria are provided, which determine to what extent a commander *may* aggregate multiple strikes/objectives into one AWS attack.

A. Contours

Drawing the line too far to the left or right of the scale is conceptually problematic in either direction. To see why, take the most restrictive option of always construing *each shot* as a separate attack. As noted above, many positions in the AWS polemics imply this proposition, but this is conceptually incorrect, inconsistent with practice and makes many related rules unworkable. Both AP I and the International Criminal Tribunal for the former Yugoslavia (ICTY) refer to '*acts* of violence'.[38] The AP I Commentary similarly mentions 'combat action' and 'counter-*attacks*',[39] implying that several individual strikes may agglomerate into one attack.[40] Bothe et al. confirm that a Shot-only interpretation was never the AP I drafters' intent.[41]

37 This is not only a logistical burden, but also precludes technologies offering benefits that cannot be achieved at human levels of performance, for example, when the time of engagement would be too short for a human response.

38 AP I (n 9) art 49(1); *Kunarac* (n 19) para 415 (emphasis added).

39 AP Commentary (n 18) para 1880 (emphasis added).

40 2013 Commentary (n 21) 329.

41 ibid.

Additionally, the proposition is inconsistent with existing practice. An attack may involve the use of force against multiple objectives that do not permit legal scrutiny per engagement.[42] Many 'defensive' systems currently in use, such as C-RAM and active protection systems,[43] would be impossible to operate if each shot were classified as an attack for which a separate legal analysis must be conducted.[44] Commanders do not supervise when each landmine 'decides' to release force or not.[45] For rocket volleys fired at a concentration of tanks, one does not look at each individual munition and determine whether it was properly directed at a military objective, whether collateral damage is excessive, etc.:[46] instead, the volley as a whole is assessed. In each of these examples, the legal analysis is performed *once*, prior to the attack's commencement.

Finally, the proportionality rule – which requires that expected collateral damage may not be excessive relative to the anticipated military advantage – precludes a Shot-only interpretation. With respect to proportionality evaluations, the possibility of counting several strikes together for the purposes of determining an attack's military advantage has been debated in the past. Commentators considered whether military advantage should be assessed on the basis of 'a single strike or for a series of strikes',[47] 'a single military objective, on the basis of a battle, a campaign or a war'.[48] Consensus was eventually reached in both scholarship and practice[49] that military advantage should 'relate to the attack considered as whole and not merely to isolated or particular parts of the attack',[50] indicating that an attack may indeed consist of multiple strikes on one or more entities.[51]

At the same time, defining 'attack' too broadly is also contrary to existing law and threatens many protections accorded by IHL. Take the proposition that '*each activation* is an attack'.[52] Depending on how long the AWS is allowed to remain active, its freedom of movement, how specific the target signature is, etc., this 'one attack' can theoretically span hours, many square kilometres, hundreds of objectives, and thousands of shots. Many protections granted by distinction and precautions would be compromised if commanders were permitted to amalgamate all this activity

42 Roff and Moyes (n 7) 5.
43 Many have been in use for several decades. See P Scharre and MC Horowitz, 'An Introduction to Autonomy in Weapon Systems' (2015) Center for a New American Security, Annex B.
44 Scharre (n 29).
45 See Moyes (n 25) 4.
46 Assuming proper corrections are made after initial shots.
47 N Durhin, 'Protecting Civilians in Urban Areas: A Military Perspective on the Application of International Humanitarian Law' (2016) 98 International Review of the Red Cross 177, 188.
48 WJ Fenrick, 'International Humanitarian Law and Combat Casualties' (2005) 21 European Journal of Population 167, 177.
49 See US Department of Defense (n 10) sct 5.6.7.3; Fleck (n 22) para 445; 2013 Commentary (n 21) 366.
50 HPCR Manual (n 20) para 1(w).
51 Y Dinstein, *The Conduct of Hostilities under the Law of International Armed Conflict* (Cambridge University Press 2016) 108.
52 Cf Scenario 1, in which case the 1200 and 1500 sorties would count as one attack respectively, for a total of two attacks.

into one attack.[53] Conceptually, it would also render the proportionality rule moot. 'Attack' can in no event stretch to encompass strikes conducted during an entire military operation[54] since this would justify almost endless levels of collateral damage on the basis that it permitted the belligerent to 'win the battle' or even the entire war.[55]

We can thus generally conclude that commanders must be allowed to conceive their AWS attacks sufficiently broadly to allow some engagements to be taken by the system in succession without human legal judgment being required for each, yet narrowly enough so as not to jeopardize the protections afforded by IHL. The question that then follows is: how do we determine the appropriate scope for *specific* AWS attacks?

B. Criteria Limiting Permissible Scope

An attack has to 'remain a finite operation with defined limits'.[56] This paper argues that two factors should inform the maximum extent to which a commander may consider successive engagements by their AWS as one attack. These are *proximity in time and proximity in space*.[57]

In AP I's *travaux préparatoires*, the International Committee of the Red Cross (ICRC) remarked that an attack 'is related to only one specific military operation, limited in space and time'.[58] Similarly, Dinstein argued that 'the temporal or geographic dimensions must be construed reasonably. They cannot be too remote or long-term.'[59] The ICTY commission investigating NATO conduct during the Kosovo War also invoked 'time or space' as a key delimiter for assessing the proportionality of particular attacks.[60] With regard to spatial proximity in particular, recall that AP I considers any attack that 'treats as a single military objective a number of *clearly separated* and distinct military objectives' as indiscriminate.[61] This presumption that time and space are decisive is reflected in works on AWS. Many commentators emphasise the need for restricting the time and space in which an AWS will operate,[62] and the Group of Governmental Experts in Geneva held that AWS users should 'limit the *duration, geographical scope*, and *scale* of the operation of the weapon system'.[63]

53 Considered from this perspective, the assertion that an AWS 'cannot proceed from one attack to another, to another, without human legal judgment being applied' makes eminent sense. See Article 36, 'Key Elements of Meaningful Human Control, Background Paper to Comments Prepared by Richard Moyes, Managing Partner, Article 36' (2016) CCW Meeting of Experts on LAWS, Geneva, 11–15 April 2016, 3.

54 Fleck (n 22) para 442.

55 L Gisel, 'The Principle of Proportionality in the Rules Governing the Conduct of Hostilities under International Humanitarian Law', (International Expert Meeting, 22–23 June 2016, ICRC 2016) 13.

56 ibid 17.

57 'Space' may need to be defined differently in the case of systems that attack intangible objectives, which fall outside the scope of this paper.

58 2013 Commentary (n 21) 329 (fn 2).

59 Dinstein (n 51) 161.

60 ICTY, 'Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign Against the Federal Republic of Yugoslavia' (2001) para 49.

61 AP I (n 9) art 51(5)(a) (emphasis added).

62 Eg Boothby (n 3) 145; P Kalmanovitz, 'Judgment, Liability and the Risks of Riskless Warfare' in Nehal Bhuta and others (eds), *Autonomous Weapons Systems* (Cambridge University Press 2016) 150.

63 GGE (n 6) para 22.

In particular, an absolute line must be drawn at the point where spatial and temporal distance starts to compromise a commander's ability to properly implement precautions in attack due to epistemic uncertainties. Since the commander

> needs to make legal judgements based on an anticipation of the interaction of a system with its operational context, there needs to be some bounding of that context in space and time in order for such judgements to be substantive. The wider the physical area, and the longer the duration of operation, the less detailed the information a commander will likely have regarding that area, and the less predictable that system's use will be.[64]

How dynamic the operational environment is, as well as the density and type of collateral concerns, influence this assessment. Some environments are relatively static or predictable (e.g., the deep sea, a demilitarized zone), but other environments in which AWS usage may be envisaged (e.g., populated areas) can change quickly and unexpectedly.[65] In complex and dynamic contexts, the maximum permitted timeframe within which objectives may still reasonably be grouped into one attack will likely be very limited, particularly if many (mobile) collateral concerns are present. Spatially, objectives grouped together in close proximity can more easily be assessed on the basis of similar surrounding circumstances, while objectives spaced further apart will inhabit distinct operational spaces that cannot be legally analysed as one attack. Finally, the commander should also consider the expected delay between engagements. There is clearly a distinction between sending multiple AWS to attack objectives simultaneously and relying on only a few (or one) AWS to locate and strike targets, since targets currently *not* being engaged may be free to move, making prior collateral estimations obsolete.

These judgments are context-specific, and commanders must carefully analyse all operational and environmental parameters to determine whether they can, in good faith, order a particular AWS attack and be reasonably convinced that all presumptions relevant for legal analysis (concerning the military nature of the objective, military advantage, collateral damage, etc.) *hold throughout* the planned autonomous attack.

One may argue that the guidelines proposed above are not specific enough or even that they are open to abuse. Would it not be better to establish more quantitative standards in terms of seconds or square kilometres? This paper argues in the negative. The extent to which objectives 'should be geographically proximate to each other, and the duration over which a use of force may constitute an individual attack, are all open questions to some extent'.[66] The aforementioned ICTY report also declined to give rigid guidelines, preferring instead to leave the question open.[67] The current author

64 Moyes (n 25) 9.
65 ICRC (n 10) para 43.
66 Roff and Moyes (n 7) 5.
67 ICTY (n 60) para 49.

views this flexibility as desirable. IHL is a practical legal regime that recognizes that each targeting situation is unique.[68] Concepts such as maximum permissible Attack Scope should, therefore, not be reduced to mathematical standards: as with precautionary obligations as a whole, some margin of discretion should be left open to allow the reasonable AWS commander to 'make a good faith judgment' regarding Attack Scope.[69] At the same time, the discussion in this section should provide a sufficient basis for superiors, legal analysts and post-hoc adjudicators to identify those situations where a commander clearly could *not* have considered that a particular choice of Attack Scope was reasonable.

4. APPLICATION TO TARGETING SCENARIOS

This section demonstrates, with the help of two scenarios, how the above methodology can be applied in practice to determine the scope of an AWS attack. This exercise also provides insights into how associated targeting obligations would be impacted.

A. How Many Attacks?

To start, let us return to Scenario 1 and attempt to answer the question posed there: how many attacks is Commander-A responsible for today? For the 1200 sortie, the tanks are co-located, and the AWS is presumably poised to engage all objectives in short succession. For this situation, it is argued that Commander-A is entitled to analyse the sortie as a single Group-attack.[70] Commander-A only needs to verify the tanks' military nature once before the engagement, and may aggregate military advantage and collateral damage, etc. This is the same way in which an artillery volley against the same group of tanks would be characterized and assessed.

For the 1500 sortie, however, the tanks are geographically separated, presumably by a significant distance. In addition, the scenario makes clear that Commander-A is only sending *one* AWS to attack all four tanks, which would entail travel time that must be taken into consideration. It is argued that this operation cannot be treated as a single Group-attack, but rather, that it must be treated as four Single-attacks. This entails separate proportionality calculations, separate attempts at mitigating collateral damage, etc. Crucially, Commander-A must continually monitor the *other* tanks while their AWS is occupied with the first, and cancel the subsequent strikes if circumstances legally demand it (e.g., if one tank rolls up to a market stall).[71] This is the same assessment that would need to be made were artillery to be used on the four

68 WB Huffman, 'Margin of Error: Potential Pitfalls of the Ruling in The Prosecutor v. Ante Gotovina' (2012) 211 Military Law Review 1, 17.

69 ibid 49.

70 Recall that for Group-attacks, the objectives must all belong to a Specific Target Group, which is the case for a tank platoon.

71 AP I (n 9) art 57(2)(b).

tanks; that is, they would constitute separate attacks and, thus, require distinct legal analyses.

This example shows how significantly Attack Scope impacts a commander's options for allowing AWS to operate autonomously. As hypothesized in Section 1, there is no simple answer as to whether autonomy is or is not allowed under IHL: it is always a question of whether said autonomy allowed is sufficiently narrow in terms of Attack Scope.

The answer to the question of how many attacks Commander-A is responsible for today is, thus, five: one Group-attack at 1200 hours and four Single-attacks at 1500.

We have now seen examples of Single- and Group-attacks. Can attacks also be of other tiers? A *Shot-attack* is a special Single-attack, where the AWS is only expected to make one engagement decision (e.g., 'kamikaze'-munitions[72]). Are offensive[73] *Signature-attacks* legally possible? While rarer, theoretically, yes—if the temporal and spatial conditions remain satisfied. Suppose a mix of enemy tanks and fuel trucks are moving through a small area, and the commander releases a swarm of AWS that can accurately identify and destroy both types of vehicles and rapidly destroy them. This would constitute a lawful Signature-level attack, but only if the fleet is sufficiently numerous to neutralize all objectives swiftly,[74] narrow geographical restrictions are applied, and there is little risk of changes in the environment during this period.

B. Assessing Proportionality

For a second demonstration that focuses more on proportionality and the duty to cancel/suspend, consider the following:

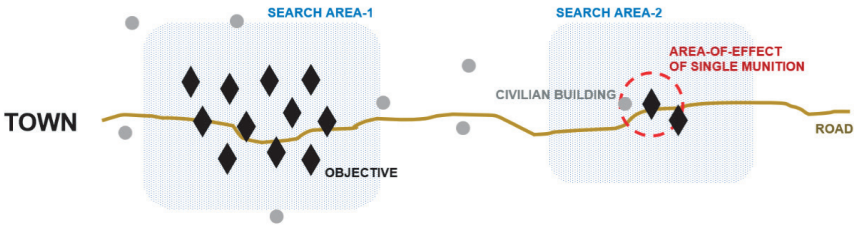
> **Scenario 2.** Enemy armour has been positively identified on a road as depicted in Figure 2. Commander-B considers sending AWS that can very reliably search and strike armour until they are neutralized, even when they take evasive action.

72 Note that if *multiple* munitions are sent (e.g. a kamikaze-swarm), the attack could potentially be qualified as a Group- or Signature-attack instead of individual Shot-attacks, depending on spatial and temporal proximity.

73 It was noted above that many defensive systems such as C-RAM and air defence fall in the Signature-tier, and are being used without controversy.

74 If only a single AWS is sent, it is likely too 'slow' since it must engage each objective in succession, and other objectives will move away, foreseeably decreasing spatial proximity.

FIGURE 2: ATTACK ON ROAD



There are no concerns regarding AP I Articles 51(4) and 57(2)(a)(i);[75] however, under current conditions, the building in Area-2 is likely to be damaged. Is this attack proportional? That depends on how we characterize this scenario in terms of Attack Scope, which admits two[76] possibilities:

(a) Two Group-attacks, comprising Area-1 and Area-2, respectively. For Area-2, one building would be damaged just to destroy two tanks, which is probably excessive.
(b) One Signature-attack, with the length of the road as the AWS's search area. One building would be damaged to destroy thirteen tanks. This is likely justifiable.

In such situations, commanders must carefully consider what Attack Scope to apply, as it significantly affects what actions are lawful under the proportionality rule. In the current scenario, whether Commander-B may count this as a single Signature-attack will likely depend on the particular circumstances. The map lacks a scale, but amalgamating the military advantage of Area-1 and Area-2 would be more justifiable if the distance between areas were 50 metres, compared to 500 metres. One can also consider the size of the fleet. If Commander-B releases multiple AWS simultaneously against Area-1 and Area-2, it would be easier to construe the scenario as a unified attack, compared to if a single AWS must 'finish' with Area-2 first before moving to Area-1. Once again, the temporal and spatial dimensions are decisive.

As theorized in Section 2, commanders will be tempted to define AWS attacks as broadly as possible because doing so carries significant practical benefits. In this example, if construed as two Group-attacks, the commander might have to delay or cancel the attack on Area-2, which may frustrate attempts to clear the road. Note that if clearing the road is *imperatively* important, then even a Group-attack on Area-2 could become justifiable, since the military advantage would derive not only from the destruction of matériel but also from the broader context (e.g., reinforcing the

75 The system is reliable enough to not be indiscriminate, and the objectives were clearly identified and validated by Commander-B.
76 Given the tanks clearly constitute two Specific Target Groups, we disregard the option of 13 Single-attacks.

town garrison that is about to collapse). Ultimately, this illustrates how judgments concerning Attack Scope (and the proportionality assessments that flow from them) are very context-dependent: commanders must utilize the guidelines presented above in good faith and not to justify *a priori* disproportionate attacks.

5. CONCLUSION

Innovations in technology constantly require us to revisit existing IHL concepts. The notion of *attack* may initially seem unproblematic with respect to AWS, but closer inspection indicates that much legal indeterminacy exists with regard to how broadly commanders can define the *scope* of AWS attacks. It was hypothesized that uncertainty on the appropriate scope of AWS attacks is a major explanatory factor of the international disagreement on the level of autonomy that AWS are legally allowed to exhibit, and this paper is one attempt to re-frame the problem through a new lens that may advance the overall debate.

To address the problem of indeterminacy, this paper proposed a scaling methodology in the form of an Attack Scope spectrum, which can help commanders to determine in more transparent terms how they characterize AWS attacks in terms of scope. Two scenarios[77] were subsequently presented and discussed to demonstrate the methodology in practice. From this analysis, it was found that an attack can technically fall in any tier within this spectrum depending on the particular circumstances of the operation, but that the appropriate Attack Scope will depend on a balance of military and humanitarian interests. The military perspective dictates that Attack Scope must not be restricted too narrowly if this would render the execution of reasonable military operations and the application of the proportionality rule unworkable. The contrasting humanitarian perspective dictates that AWS attacks must be restricted in time and space so as not to jeopardize the protections that IHL grants to the civilian population through the principles of discrimination and proportionality. Ultimately, the guidelines presented in this paper sacrifice neither perspective – an approach that reflects the core balancing philosophy of IHL.[78]

Greater clarity and transparency concerning Attack Scope are liable to positively affect both the belligerent and the civilian. The AWS commander is strongly encouraged to consciously consider Attack Scope before launching any AWS attack. Knowing at which intervals they are imperatively required to perform legal assessments during their AWS's operational cycle removes legal ambiguity. Additionally, in cases where

77 The two scenarios presented in this work were designed to illustrate the concepts introduced in this paper in practical conditions. Future work may involve the study of more variations of toy scenarios to theorize, for example, if there are generally accepted limits in space and time which most reasonable commanders would consider as absolute boundaries with regard to stretching Attack Scope, which may further refine the general guidelines offered in Section 3.

78 AP Commentary (n 18) para 1389.

they are later called to account for any (incidental) harm caused by their AWS,[79] such reflection will allow them to justify how they delineated the contours of their attack(s) and explain for each attack why they did not deem the collateral damage to be excessive. For the civilian, clear boundaries to the extent to which commanders may extend the notion of attack positively impact their protection under IHL, in particular by requiring commanders to execute (re)validation and (re)assessment of proportionality at appropriate moments and by clarifying when commanders must maintain oversight over the *next* attack when AWS are used for successive engagements.

79 US Department of Defense (n 10) sct 5.10.2.2.

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

Targeting in the Black Box

Scott Sullivan*
Professor
Army Cyber Institute
United States Military Academy
West Point, NY, United States
scott.sullivan@westpoint.edu

Iben Ricket
Research Scientist
Dartmouth College
Hanover, NH, United States
iben.sullivan@dartmouth.edu

Abstract: Artificial intelligence (AI) is poised to become pervasive in military operations worldwide. In the coming decades, AI-based systems will revolutionize logistics, dramatically change targeting, and ultimately power autonomous weapons systems. Unfortunately, many of the most potent AI-based systems are unintelligible to their developers and offer unexplained outputs to their users—a phenomenon called the "black-box problem." This paper first describes the basic AI architecture that is giving rise to the black-box problem. It then shifts to consider the unexplored question of whether black-box models comport with the international humanitarian law (IHL) principles of distinction, proportionality, and precaution, which are fundamentally rooted in nuanced context and subjective judgment. After describing the mismatch between black-box models and existing IHL principles, the paper compares existing NATO doctrine with emerging "soft law" embraced by NATO member States. Identifying a nascent movement away from explainable AI, the paper concludes by setting out the importance of interpretability and the aspects therein that should be considered by policymakers in constructing future legal norms and by military officials in assessing AI models to be used in future operations.

Keywords: *AI, targeting, autonomous weapons, Israel, neural networks, explainable AI*

* The views expressed in this article are personal and do not reflect the policy or position of any US government entity or organization.

1. INTRODUCTION

At the end of 2023, multiple outlets reported on Israel's widespread use of an artificial intelligence (AI) system to identify targets in its ongoing conflict with Hamas in Gaza.[1] Dubbed "Habsora," or "the Gospel" in English, Israeli Defense Forces (IDF) officials credited the AI system with the ability to increase the number of targetable sites in Gaza from 50 each year to over 100 each day.[2] Commentators quickly recognized AI-based targeting as "an intermediate step [to] autonomous systems that will eventually be deployed to the battlefield."[3]

While various forms of AI targeting and autonomous weapons systems have long been in service, their use has primarily been restricted to circumstances where the risk to civilians was minimal and the targeting question at issue was easy to identify definitively.[4] The introduction of more generally purposed AI-based targeting systems, like Habsora, in service of a conflict against a terrorist organization based in one of the most densely populated areas of the world marks a definitive shift.

AI systems that can target and use force without human intervention, often called lethal autonomous weapons systems (LAWS), have engendered the greatest attention and concern. Proponents suggest that well-designed LAWS may operate more humanely and lawfully by avoiding classically human errors and cognitive biases. Skeptics believe LAWS usage will generally increase armed conflict and risk unpredictable and perhaps unknowable dangers to combatants, civilians, and the larger global order. However, all parties acknowledge that the development and deployment of such systems are inevitable, leading to a recent call to "establish specific restrictions on autonomous weapons systems" by the leaders of the United Nations and the International Committee of the Red Cross (ICRC).[5]

Unfortunately, the attention garnered by the specter of robot soldiers tends to obscure the fact that AI systems powering autonomous weapons, and not the execution of force itself, pose the most imminent and widespread challenge for IHL regulation. Targeting systems like Hasbora exemplify this concern. Much of the system's operations remain

1 *See e.g.*, Harry Davies, Bethan McKernan & Dan Sabbagh, *"The Gospel": How Israel Uses AI to Select Bombing Targets in Gaza*, Guardian (UK) (Dec. 1, 2023), https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets.

2 *Id*. The United States has also recently begun using an AI-based targeting system for operations against Yemen's Houthis and Iran proxy forces. Katrina Manson, *AI Warfare Is Already Here*, Bloomberg (Feb. 28, 2024), https://www.bloomberg.com/features/2024-ai-warfare-project-maven/.

3 Geoff Brumfiel, *Israel Is Using an AI System to Find Targets in Gaza. Experts Say It's Just the Start*, Morning Edition, NPR (Dec. 14, 2023), https://www.npr.org/2023/12/14/1218643254/israel-is-using-an-ai-system-to-find-targets-in-gaza-experts-say-its-just-the-st.

4 Jack M. Beard, *Autonomous Weapons and Human Responsibilities*, 45 Geo. J. Int'l L. 617, 628–32 (2014).

5 *Joint Call by the United Nations Secretary-General and the President of the International Committee of the Red Cross for States to Establish New Prohibitions and Restrictions on Autonomous Weapon Systems*, International Committee of the Red Cross (Oct. 5, 2023), https://www.icrc.org/en/document/joint-call-un-and-icrc-establish-prohibitions-and-restrictions-autonomous-weapons-systems.

out of view. However, its apparent discovery of hundreds of "new" targets, coupled with a large and growing civilian death toll and the widespread destruction of civilian objects, has engendered tremendous criticism of Israeli targeting decisions and its dependence on AI.[6]

This paper seeks to build on the existing literature regarding autonomous weapons to discuss the legal and policy implications surrounding AI modeling systems, such as artificial neural networks, which are poised to fuel the targeting and autonomous capabilities of the future. These AI systems, often referred to as "black-box models," function unintelligibly to the States that develop them and generate outputs that offer little, if any, underlying reasoning to the personnel operating them.

Black-box models, such as artificial neural networks, sometimes called "deep learning," are widely considered the most powerful and accurate AI prediction systems on offer. However, their unintelligible operations and unexplained outcomes stand in contrast to prior iterations of machine learning methods and the software powering earlier "automated" or "autonomous" weapons, which function through increasingly complex, albeit explainable technology. Consisting of multiple layers of interconnected nodes, artificial neural networks require enormous volumes of data and produce outcomes or classifications reflecting "extremely complex non-linear statistical models and innumerable parameters" that lack any "reason or suitable explanation" discernible to the user, or even the architects of the system.[7]

The use of black-box models in targeting poses significant practical and legal challenges under established principles of IHL, a body of law whose principles are built upon a bedrock of context and subjectivity.

2. THE EXPLAINABILITY PROBLEM IN AI

Inherent to all AI systems is the ability to learn patterns in the data to generate a prediction.[8] The nature of the relevant patterns and predictions, of course, depends upon the field. AI models in medicine can identify patients at risk for hospital

6 Evan Hill, Imogen Piper, Meg Kelly & Jarrett Ley, *Israel Has Waged One of This Century's Most Destructive Wars in Gaza*, Washington Post (Dec. 23, 2023), https://www.washingtonpost.com/investigations/interactive/2023/israel-war-destruction-gaza-record-pace/; Jared Malsin & Saeed Shah, *The Ruined Landscape of Gaza After Nearly Three Months of Bombing*, Wall Street Journal (Dec. 30, 2023), https://www.wsj.com/world/middle-east/gaza-destruction-bombing-israel-aa528542.

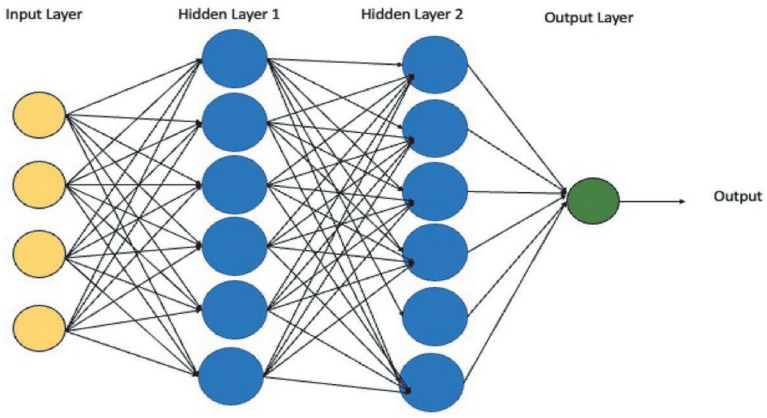
7 Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud & Amir Hussain, *Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence*, 16 Cognitive Computation 45 (2023).

8 Iben Sullivan (Ricket), Michael Matheny & Jeremiah Brown, *Breaking Down the "Black-box" of Machine Learning for Predictive Analytics: Results from Models Predicting 30-Day Readmissions Following an AMI*, Nurs. Sci. Conf. Proc. (2023) (on file with the authors).

readmissions or predict the presence of early cancer.[9] Credit card companies and banks use AI algorithms in anomaly detection to identify fraudulent or abusive transactions. Autonomous vehicles rely on AI systems trained on data to identify appropriate and safe driving performance.

While different AI modeling techniques present varying degrees of explainability problems, this paper focuses on neural networks and related deep-learning-based models, which are widely considered the most effective in classification and evaluation and also among the most opaque of AI systems.[10]

FIGURE 1: BASIC NEURAL NETWORK ARCHITECTURE. THE NEURAL NETWORK INCLUDES AN INPUT LAYER WHERE DATA IS FED INTO THE ALGORITHM, TWO HIDDEN LAYERS, AND AN OUTPUT LAYER WHERE DATA IS SCALED TO AN APPROPRIATE RANGE USING AN ACTIVATION FUNCTION



At their most basic structure, as shown in Figure 1, neural networks contain an input layer, one or more hidden layers, and an output layer.[11] Each layer includes nodes or neurons.[12] Data is passed through each layer via a neuron. Each neuron can be considered its own model, with input data, weights, a threshold (bias), and an output.[13] Each input is assigned a weight with larger values signifying greater importance. All inputs are multiplied by their respective weights, summed, and passed through an activation function.[14] If this final output exceeds a specific threshold, the

9 Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis & Dimitrios I. Fotiadis, *Machine Learning Applications in Cancer Prognosis and Prediction*, 13 Computational and Struct. Biotech. J. 8, 8 (2015).

10 Manuel Carabantes, *Black-Box Artificial Intelligence: An Epistemological and Critical Analysis*, 35 AI & Soc. 309, 313 (2020).

11 Rene Y. Choi, Aaron S. Coyner, Jayashree Kalpathy-Cramer, Michael F. Chiang & J. Peter Campbell, *Introduction to Machine Learning, Neural Networks, and Deep Learning*, 9 Transl. Vis. Sci. Technol. no. 2, art. 14 (2020).

12 *Id.*

13 Bogdan M. Wilamowski, *Neural Network Architectures and Learning Algorithms*, 3 IEEE Indust. Electr. Mag. no. 4, 2009, at 56–63.

14 *Id.*

neuron is activated, and the data is passed to the next network layer. As such, one neuron's output is the next layer's input.[15] Data is passed from one layer to the next until it reaches the final output layer, which defines the acceptable range of output (e.g., -1–1 or 0–1) by scaling or transformation.[16] While simple feedforward neural networks may contain a few hidden layers, where data moves through the network in one direction (forward), more complex and deeper networks can contain hundreds or thousands of hidden layers and feed data forward and backward through the layers.[17] Special neural network architectures exist for specific data inputs or use cases.[18] For example, convolutional neural networks are designed for images, and recurrent neural networks are used for text.[19]

The architecture and processes inherent to neural networks pose two fundamental problems of understandability. The first relates to complexity. Even a cursory explanation of neural network architecture illustrates the intractable challenge of following inputs through the network to a final predicted output.[20] As the complexity of the system increases, its opacity rises as well.[21] The most accurate and powerful neural network models—presumably the type of models States would desire to build and use in armed conflict—require endless volumes of data and create millions, if not billions, of parameters.[22] Thus, a single prediction generated from a deep learning algorithm can involve millions of mathematical and computational operations, making traceability from data to prediction (or back) functionally impossible.[23]

Furthermore, the nature of a neural network's architecture and operations, regardless of type, size, or complexity, is so cognitively dissimilar to humans that there exists a fundamental "mismatch between [the model's] nature and [human] understanding," and as a result, the models "are not intelligible no matter how much knowledge we possess on mathematics, computation, or any other related science."[24]

Collectively, these challenges for users and developers to understand how certain AI models operate and formulate specific predictions are commonly called the "black-box problem."[25] The use of black-box models has proliferated in recent years in areas of

15 Choi et al., *supra* note 11, at 14.
16 Wilamowski, *supra* note 13, at 56–63.
17 Choi et al., *supra* note 11, at 14.
18 *Id*.
19 *Id*.
20 Plamen Angelov & Eduardo Soares, *Towards Explainable Deep Neural Networks (xDNN)*, 130 Neural Netw. 185, 185–89 (2020).
21 Choi et al., *supra* note 11, at 20.
22 Angelov & Soares, *supra* note 20, at 188.
23 *Id*.; Rabia Saleem, Bo Yuan, Faith Kurugollu, Ashiq Anjum & Lu Liu, *Explaining Deep Neural Networks: A Survey on the Global Interpretation Methods*, 513 Neurocomputing 165, 165–68 (2022).
24 Carabantes, *supra* note 10, at 314.
25 Pantelis Linardatos, Vasilis Papastefanopoulos & Sotiris Kotsiantis, *Explainable AI: A Review of Machine Learning Interpretability Methods*, 23 Entropy, no. 1, 2020, at 3; Scott M. Lundberg et. al., *From Local Explanations to Global Understanding with Explainable AI for Trees*. 2 Nature Mach. Intell. 56 (2020).

decision-making with high-stakes outcomes, such as healthcare and criminal justice.[26] In such areas, the perceived higher performance of black-box models relative to their simpler counterparts is pitted against the inscrutability—and potential hidden biases and errors—intrinsic to black-box AI.[27]

3. EXISTING LAW AND THE ROLE OF EXPLANATIONS

Explanations are fundamental to functioning legal systems. In domestic legal systems, law enforcement officers must be able to articulate the evidence justifying their arrests. Judges author opinions to explain how they arrived at their decisions and to further define the contours of legal rules for future controversies. Administrative agencies are required to explain the reasoning behind their establishment of regulations.

Explainability may be even more significant in the international legal system. International legal obligations are typically drafted at a much higher level of abstraction than their domestic counterparts due to their subjects' tremendous economic, cultural, and legal differences.[28] Despite this, the absence of a "centralized enforcement mechanism" to coerce compliance means that the force of international legal rules flows, at least in part, on States' ability to understand the boundaries of the rule and the reasoning underlying such boundaries.[29]

Within the specific context of IHL, the legality of State action frequently turns on the veracity of the proffered explanation. Such explanations can, for instance, serve to distinguish between an accepted act of war and a war crime. However, this paper focuses on how the larger structural principles of IHL embed the concept of explainability of targeting decisions and how reliance on black-box AI systems threatens the fabric of that system.

There is general agreement that AI weapons systems, autonomous or not, must conform with existing IHL. NATO doctrine provides that AI weapons systems must be "developed and used in accordance with … international humanitarian law and human rights law."[30] The more recent US-led *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy* similarly stipulates that the "use of AI in armed conflict must be in accord with States' obligations under international

26 *See* Cynthia Rudin, *Stop Explaining Black-Box Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 Nature Mach. Intell. 206, 207–10 (2019).

27 *Id*. The exchange between a model's inherent understandability and its performance is often referred to as the "performance–interpretability tradeoff."

28 *See* Jack Goldsmith & Daryl Levinson, *Law for States: International Law, Constitutional Law, Public Law*, 122 Harv. L. Rev. 1791, 1824 (2009).

29 *Id*.

30 Zoe Stanley-Lockman & Edward Hunter Christie, *An Artificial Intelligence Strategy for NATO*, NATO Review (Oct. 25, 2021), https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html.

humanitarian law including its fundamental principles."[31] One of the significant challenges with assessing the legal status of AI on the battlefield, especially in complex combat circumstances, arises from the fact that the applications and architecture of such systems are largely theoretical or legally protected secrets. As legal scholars argued nearly a decade ago, "there is no reason, in principle, why a highly automated or autonomous system could not satisfy the requirements of targeting law."[32] The same could be said of an AI-based targeting system.

With the dawn of AI-based targeting upon us, our attention must turn to the designers of these systems and the commanders employing them, who will be required to exercise judgment in a manner consistent with IHL. Of course, responsible judgment can only occur when it is well-informed. It is this informed judgment requirement, which is embedded in existing humanitarian law and the emerging norms surrounding AI weapons systems, that black-box AI models challenge.

A. The Principles of Distinction and Proportionality

Distinction and proportionality reflect two of the most basic legal principles underlying IHL.[33] Distinction is often considered the principle upon which AI-based weapons systems are most likely to succeed, especially if success is judged by their perceived ability to outperform their human counterparts. Distinction requires a combatant to use "reasonable judgment" to differentiate combatants and military objects from civilians and civilian objects.[34] A variety of peculiarly human traits with tragic consequences often compromises human targeting decisions.[35] Untouched by these factors, machine-learning systems armed with vast quantities of high-quality data gathered over time would presumably be well-positioned to identify the patterns of combatants and military objects and distinguish them from civilians and civilian objects. Such a system would be especially effective where, such as in the war in Ukraine, most combatants would be targetable based on their status as members of a State's armed forces or another organized armed group.

Legal nuances quickly proliferate when considering the targeting of civilians. Civilians are generally protected from attack but forfeit that protection when "directly

31 *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy*, U.S. Dep't of State (Nov. 9, 2023), https://www.state.gov/wp-content/uploads/2023/10/Latest-Version-Political-Declaration-on-Responsible-Military-Use-of-AI-and-Autonomy.pdf.

32 Kenneth Anderson, Daniel Reisner & Matthew Waxman, *Adapting the Law of Armed Conflict to Autonomous Weapons Systems*, 90 US Int'l L. Stud. 386, 406 (2014).

33 *Id*. at 401.

34 Geoffrey S. Corn, *Targeting, Command Judgment, and a Proposed Quantum of Information Component: A Fourth Amendment Lesson in Contextual Reasonableness*, 77 Brook. L. Rev. 437, 454 (2012) ("each ad hoc targeting decision must be the result of a reasonable judgment").

35 *See generally*, Kevin Jon Heller, *The Concept of "the Human" in the Critique of Autonomous Weapons*, 15 Harv. Nat'l Sec. J. 1 (2023).

participating" in hostilities.[36] While many instances of direct participation are uncontroversial—for example, firing upon combatants—the scope of what behaviors constitute direct participation is undefined and largely undefinable. For example, *The Commander's Handbook on the Law of Naval Operations*, issued by the US, indicates that "there is no definition of direct part in hostilities in international law" and, as such, combatants "must make an honest determination" based on "all relevant available facts in the circumstances prevailing at the time."[37] Even when a civilian might be found to be directly participating, a temporal question arises. Because civilians are only targetable "so long as" they directly participate in hostilities, their targetability ceases when their participation ceases. This temporal aspect has led to competing standards, both of which require a nuanced, case-by-case assessment that is also highly transient, given the innumerable ways by which participation might cease.[38]

Despite these intricacies, the determination emanating from a black-box AI system would ultimately be binary, with each person or object delineated as either a target or non-target. Beneath the binary nature of the classification would be a probabilistic determination of the model's prediction. In the absence of any explanation from the system regarding how the model came to its conclusion, there is little "judgment" to be exercised by the commander who receives the model's determination. He is either to trust the system or not. The "honest judgment" of the commander is replaced by the decision to trust the AI system or not.[39]

Proportionality analysis poses far more substantial challenges. The principle of proportionality demands that "the incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof" of an attack not "be excessive in relation to the concrete and direct military advantage" to be accomplished.[40] In essence, proportionality requires the balancing of two qualitatively different interests. Of these, insofar as the reliable identification of civilian persons and objects can be

36 Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts, June 8, 1977 [hereinafter AP I], art. 51; *see also* Nils Melzer, Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law (International Committee of the Red Cross 2009), https://www.icrc.org/en/doc/assets/files/other/icrc-002-0990.pdf.

37 Dep't of the Navy, The Commander's Handbook on the Law of Naval Operations, NWP 1-14M/MCTP 11-10B/COMDTPUB P5800.7A, § 8.2.2 (2022). An earlier iteration of the handbook similarly stated that "direct participation in hostilities must be judged on a case-by-case basis" as to each "particular civilian" that might be subject to attack. Dep't of the Navy, The Commander's Handbook on the Law of Naval Operations, NWP 1-14M, § 8.2.2 (2007).

38 The "continuous combat function" espoused by the ICRC requires a fine-grained analysis of acts that give rise to a quantitative and qualitative assessment of when sporadic involvement in hostilities passes the relevant threshold. *See* Melzer, Interpretive Guidance, *supra* note 36, at 33–6. In contrast, the membership test poses a similarly intensive examination of whether an individual's social network and behaviors are sufficiently tied to an armed group so as to render him targetable as a member of that group.

39 While the prudential considerations attached to exercising judgment on whether to attack a targetable person or object are not strictly required by distinction, the practical limitations on a commander's options are significant.

40 Customary International Humanitarian Law (Jean-Marie Henckaerts & Louise Doswald-Beck eds., International Committee of the Red Cross 2005), Rule 14.

achieved, identifying the anticipated cost of an attack might be the easiest aspect for an AI-based targeting system to accomplish. To the extent that a system can correctly identify civilians and civilian objects, their presence (or absence) combined with the means of the attack potentially used in the operation would pose a relatively straightforward assessment of civilian cost. In contrast, the variations on the "concrete and direct military advantage" of a specific attack are much more subjective and, as such, difficult to capture by quantitative means. Put simply, "measuring concrete and direct military advantage will always be part of subjectivity" as it weighs elements "that are not quantifiable."[41]

Understanding an operation's direct and concrete military advantage requires an appreciation of how a specific attack fits within a larger operation and the tactical and strategic advantages the attacking party seeks to capture. Consequently, the quantitative and qualitative aspects of military advantage are not only highly nuanced—similar to distinction—but also highly dynamic and fluid. For example, the military advantage of destroying an arms depot is much higher when new intelligence has identified it as the only repository of specific munitions upon which your enemy relies. States and scholars alike disagree on the scope of considerations relevant to assessing the "concrete and direct military advantage" of a target and whether the assessment should be limited to individual attacks or extended to a broader scale.[42]

While States have proposed different formulations of engaging in proportionality analysis, there is consensus that human judgment governs the analysis. The German government, for instance, has stated that the relevant calculations must be made on "a case-by-case basis and that no abstract calculations [are] possible."[43] Canada requires its commanders to possess a subjective "honest and reasonable expectation that the attack will make a relevant contribution to the success of the overall operation."[44] Further, the gravity of the military advantage includes the "security of the attacking forces."[45]

AI targeting platforms like Habsora undermine rather than effectuate the subjective standards enshrined in distinction and proportionality. Israel, like other States, has explicitly rejected the existence of a uniform calculation for balancing concrete military advantage in proportionality analysis.[46] However, descriptions of how Habsora communicates proportionality judgments to the units carrying out attacks

41 AP I, *supra* note 36, art. 51(5).

42 *See* Nelleke H. Hoffs, *Deducing the Measuring Standard of "Concrete and Direct Military Advantage Anticipated," Referred to in Article 51(5)(b) of Additional Protocol I to the 1949 Geneva Conventions* (Sep. 21, 2013) at 8, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2329212.

43 *See ICRC—Customary IHL—Practice Relating to Rule 14 Proportionality in Attack*, Government of Germany https://ihl-databases.icrc.org/en/customary-ihl/v2/rule14?country=de.

44 Office of the Judge Advocate General, Canada, The Law of Armed Conflict at the Operational and Tactical Level, p. 4-3, §§ 20 and 21 (1999).

45 *Id*. Presumably the more insecure the attacking forces, the lower the threshold of the advantage requirement.

46 IDF Sch. of Military Law, Isr., Rules of Warfare on the Battlefield 27 (2nd ed. 2006).

do not include the subjective balancing requirement the law requires. Instead, targets displayed to commanders are accompanied by either a "green, yellow, [or] red [light], like a traffic signal."[47] The details of how Habsora determines when the red light turns to yellow or green are unknown to the public and apparently unknown to the IDF units relying on these outputs. However, the interface itself suggests a simplicity that belies the complicated analysis accompanying proportionality and encourages human deference to the algorithmic determination that the target is a "go" for a strike.

The "case-by-case" emphasis in assessing civilian "direct participation" and proportionality in attacks is deliberate and reflects the considered judgment of IHL that such legal principles are intended to comport with standards rather than rigid rules. Standards "rely on case-by-case decisionmaking and render a decision that is ostensibly limited to the facts of the case."[48] Standards broaden discretion and, consistent with the distinction and proportionality described above, do so based on the subjective assessment of a commander in the field's evaluation of the distinctive set of facts. As a result, standards "allow for the decrease of errors of underinclusiveness and overinclusiveness" and "allow the decisionmaker to take into account all relevant factors or the totality of the circumstances."[49]

In short, standards like "direct participation" and proportionality analysis exist because of the judgment that the amount of factual variance is near infinite, information is imperfect, and the error costs (both to commanders and civilians) demand each situation to be judged individually. Such circumstances are a poor fit for AI systems dependent upon enormous volumes of high-quality data and unable to absorb qualitative, subjective content such as command intentions that "cannot be automated with narrow [i.e., non-AGI] AI technology."[50] Worse, their use, especially when unaccompanied by an understanding of the variables upon which the system came to its output, is more likely to impair rather than augment the human judgment upon which these complex legal standards rely.

B. The Principle of Precaution and "Constant Care"

At its core, the precautionary principle is an obligation effectuated in planning. The obligation, reflected in customary law and Article 57 of Additional Protocol I, creates a general obligation applicable to all military operations and imposes specific requirements regarding attacks. Specifically, Article 57(1) imposes an affirmative obligation on States to take "constant care" in the "conduct of military operations" to spare civilians and civilian objects.[51]

47 Davies, McKernan & Sabbagh, *supra* note 1.
48 Edward Lee, *Rules and Standards for Cyberspace*, 77 Notre Dame L. Rev. 1275, 1295 (2002).
49 Kathleen M. Sullivan, *The Supreme Court, 1991 Term—Foreword: The Justices of Rules and Standards*, 106 Harv. L. Rev. 22, 58–9 (1992).
50 Avi Goldfarb & Jon R. Lindsay, *Prediction and Judgement: Why Artificial Intelligence Increases the Importance of Humans in War*, 46 Int'l Sec. no. 3, Winter 2021/22, at 9.
51 AP I, *supra* note 36, art. 57(1).

Article 57(2) imposes more specific precautions "with respect to attacks."[52] Under this provision, commanders must "do everything feasible to verify that the objectives to be attacked are neither civilians nor civilian objects" and to refrain from imposing civilian damage disproportionate to the anticipated military advantage.[53] Further, States are obligated to suspend any attack "if it becomes apparent that the objective is not a military one," provide "effective advance warning" of attacks when possible, and choose objectives causing lesser civilian damage when such choice is available "for obtaining similar military advantage."[54]

The inexplicability of black-box AI targeting systems poses fundamental problems under both the general duty of constant care of Article 57(1) and the precautions in attack obligations set out in Article 57(2).

First, the more specific precautionary obligations set out in Article 57(2) are dependent upon commanders understanding the basis of the targeting recommendation made by any AI-based system. For instance, the ability to verify a target as required by Article 57 can only occur if those tasked with verification are aware of the information giving rise to the targetability determination and how it was weighed. Similarly, awareness of the factors an AI model is basing its determination on might influence a commander's decision to engage in an alternative means of attack.

The broader "constant care" obligation imposed under Article 57(1) has implications for a State's obligations in attack as well as in its preceding design of targeting platforms. The duty of constant care requires States to possess "situational awareness at all times" in identifying and averting avoidable civilian harms associated with targeting decisions.[55] This includes understanding the variables giving rise to initial determinations of targetability (or non-targetability), the military advantages perceived, the civilian harm envisioned, *and* the recognition that such conditions are dynamic and that newly available information might alter the relevant legal calculus. In short, such situational awareness requires an understanding of the AI-based targeting model at a global level (how it operates generally) and at the local level (why it generated specific outputs)—neither of which is available via black-box models.

Moreover, the broad scope of the duty of constant care requires States to ensure explainability as a matter of design, not just at the time of the use of armed force. As the language suggests, the "constant care" obligation, unlike the distinction and proportionality requirements, extends beyond armed attacks and is "relevant in both

52 *Id*. art. 57(2).
53 *Id*.
54 AP I, *supra* note 36, art. 57(2), 57(3).
55 Eric Talbot Jensen, *Cyber Attacks: Proportionality and Precautions in Attack*, 89 Int'l L. Stud. 198, 202 (2013).

peacetime and *wartime*."[56] According to the ICRC, military operations covered under the duty include "any movements, manoeuvres and other activities whatsoever carried out by the armed forces with a view to combat."[57] As explained by Russell Buchan, an "operation possesses a military character where it is designed to advance combat."[58] As a result, there is reason to believe that the development and deployment of AI targeting systems would constitute "military operations" subject to the State's "constant care" obligations.[59] Just as the duty of constant care can be violated by a commander failing to weigh the relevant factors in executing an attack, States can violate the duty of constant care by designing and deploying a black-box targeting system that systematically precludes commanders from performing effective due diligence on suggested attacks.

4. BEYOND EXISTING LAW

There is consensus that existing legal principles are insufficient to govern the use of AI systems in armed conflict. Unfortunately, there is little agreement about the contours of new legal norms governing such systems. The ambiguity of what norms should govern AI systems includes a lack of clarity on the importance of avoiding black-box AI and the type of explainability that should be required. The Artificial Intelligence Strategy adopted by NATO identifies "explainability and traceability" as one of its principles and states that "AI applications will be appropriately understandable and transparent."[60] More recently, the *Political Declaration on Responsible Military Use of Artificial Intelligence* espouses a narrower understandability requirement and foregoes any reference to "explainability." Regarding development, the declaration states that military AI systems should be "developed with methodologies, data sources, design procedures and documentation that are transparent to and auditable by their relevant defense personnel."[61] While reprising the "transparency" language set out in NATO doctrine, the language only addresses the transparency of the constitutive components of the model rather than the model itself. In other words, personnel are to have awareness of (and presumably understand) the fundamental elements upon which a military AI system is based but such understanding will not, in and of itself,

56 Asaf Lubin, *Lieber Studies Big Data Volume—Algorithms of Care: Military AI, Digital Rights, and the Duty of Constant Care*, Articles of War (Feb. 13, 2024), https://lieber.westpoint.edu/algorithms-care-military-ai-digital-rights-duty-constant-care/; *see also* Jensen, *Cyber Attacks, supra* note 55, at 202; Russell Buchan, Data Protection in War (2023) (on file with author).

57 Int'l Comm. of the Red Cross (ICRC), Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949, at ¶ 2191 (Yves Sandoz et al. eds., 1987).

58 Buchan, *supra* note 56, at *10.

59 There is substantial ambiguity (and thus debate) regarding where the duty of constant care begins and ends. *See e.g.*, Michael Schmitt, *Big Data: International Law Issues During Armed Conflict, in* Big Data and Armed Conflict (Laura A. Dickinson & Edward W. Berg eds., 2023). However, because inexplicability is persistently intrinsic to black-box AI models, their usage would necessarily involve contexts where the duty would undoubtedly apply.

60 Stanley-Lockman & Christie, *supra* note 30, at C.

61 *Political Declaration, supra* note 31.

guarantee any understanding of how the model is generating specific outputs or even much insight into its general operation.

The declaration also states that users should "be trained so they sufficiently understand the capabilities and limitations" of each system.[62] As with the language of "transparency," this requirement, while appropriate, does little to address the core black-box problem—the ignorance of how specific outputs are made. For example, knowing that an AI targeting platform sometimes might mistake a video camera for a weapon tells you little about whether that particular error might have occurred in any specific targeting decision.

Beyond legal obligations, an intractable concern attaches to black-box AI systems that counsel in favor of explainable alternatives: their resistance to adjustment and the hidden errors and biases their opacity may perpetuate. A limited understanding of how an AI system makes its prediction poses an inherent difficulty in identifying when it has made errors.[63] Such hidden errors would be especially challenging to detect in targeting circumstances where civilians and combatants are difficult to distinguish. When an AI system incorrectly identifies a civilian as a target, absent incontrovertible evidence to the contrary, the system user will likely characterize the subsequent strike as a success, creating a self-fulfilling prophecy problem, in which the reaction to the prediction effectively renders the prediction accurate. Just as problematically, black-box models are notoriously difficult to debug when inevitable prediction errors are discovered.[64] Imagine an AI system that targets an ambulance by mistake. Without knowing how the model came to its targeting conclusion, it is impossible to know how to ensure that such an error does not recur. The problem could relate to inappropriate weighting, missing parameters, or nuances within the original training data. The difficulty in detecting and remediating prediction errors is precisely why black-box models are considered especially vulnerable to security intrusions.[65]

An "appropriately" explainable and understood AI system depends on the domain and application in which it will be used. In the context of IHL, the explainability ought to be understood within the context of the aspects of judgment described above as embedded in the core principles of distinction, proportionality, and precaution discussed above and the balance between humanitarian interests and military necessity embedded in the regime. To that end, States need to emphasize interpretability as a necessary component of any AI-based targeting system—whether operating autonomously or not.

62 *Id*, at para. 6.
63 *See Carabantes, supra* note 10.
64 Thomas P. Quinn, Stephan Jacobs, Manisha Senadeera, Vuong Le & Simon Coghlan, *The Three Ghosts of Medical AI: Can the Black-Box Present Deliver?* 124 A.I. in Medic. 1, 3 (2022).
65 *Id*.

Model interpretability gauges the degree to which the user can understand how an AI system came to its prediction. Globally, interpretability ensures those using the AI system can verify that it performs as expected.[66] At this level, it is essential for users to understand the magnitude and direction of the input variables' impact on the final predicted outcome, enabling them to grasp, on average, the relationship between input data and predicted outcomes.[67] As to specific predictions, the user needs to understand how the model's individual features influence the final output and its predicted value.[68] This enables users to identify in detail how the value of each input combines to generate the predicted output for an individual (or single unit).

Model interpretability is also essential for understanding the mechanisms of the algorithm.[69] This becomes especially important when monitoring systems for bias. Breaking the algorithm into discrete steps illustrates the decision process.[70] This allows users to review each step the model took between input and final prediction. In reviewing the key steps, users can identify specific variables the model used to make decisions and any associated critical value it used to make the decision.

5. CONCLUSION

The growing capabilities of AI have provoked tremendous excitement and consternation. Nowhere are the stakes of this technology higher than in international security and armed conflict. However, amid the inevitable AI race gripping the world, it is crucial to remember that the legal rules underlying our interactions, both internationally and personally, were written from an unmistakably human perspective. The subjective standards and context-driven demands of IHL place human judgment at the center of authority. Relying on human subjectivity, these rules require AI systems that enable individuals to exercise their judgment in a manner consistent with existing legal constraints.

66 Franck Jaotombo, Luca Adorni, Badih Ghattas & Laurent Boyer, *Finding the Best Trade-off Between Performance and Interpretability in Predicting Hospital Length of Stay Using Structured and Unstructured Data*, 18 PLoS One, no. 11, 2023; Linardatos, Papastefanopoulos & Kotsiantis, *supra* note 25; Lundberg et al., *supra* note 25, at 56–67.

67 *Id*.

68 *Id*.

69 *Id*.; *see also* Linardatos, Papastefanopoulos & Kotsiantis, *supra* note 25.

70 *Id*.

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

The International Legal Framework for Hunt Forward and the Case for Collective Countermeasures

Jeff Kosseff*
Associate Professor
Cyber Science Department
United States Naval Academy
Annapolis, MD, United States
kosseff@usna.edu

Abstract: United States Cyber Command's "persistent engagement" strategy and "defend forward" operational concept have produced the "Hunt Forward" operation. As Cyber Command describes them, Hunt Forward operations are "strictly defensive operations" that Cyber Command conducts "at the request of partner nations." Hunt Forward protects both US allies and the United States by blunting the harm of malicious attacks on shared networks, and it provides the United States with valuable intelligence about adversaries' methods. Cyber Command has publicly reported successful Hunt Forward operations in Ukraine, Latvia, Albania, Estonia, and other nations.

This paper draws on publicly available sources, including Cyber Command reports and media commentary, to give a comprehensive picture of Hunt Forward's capabilities, operations, and limitations. The paper argues that Hunt Forward has already resulted in numerous successful operations around the world and benefited both the United States and its allies. The paper then analyzes the basis for Hunt Forward under international law and concludes that current operations, as publicly described, are permissible. The paper goes on to argue that although Hunt Forward is purely defensive, future collaborative operations should include assistance in degrading adversaries' ability to conduct malicious cyber campaigns against the United States and its allies. To provide further breathing space for collaborative operations, the global community should affirm the use of collective countermeasures, a concept that some countries, such as

* The views expressed in this paper are only the author's and do not represent the United States Naval Academy, Department of Navy, or Department of Defense.

Estonia and Costa Rica, have embraced and that others, such as Canada and France, have questioned.

Keywords: *countermeasures, sovereignty, Hunt Forward, retorsion, espionage*

1. INTRODUCTION

Over the past decade, US cyber strategy has evolved to address new threats, gradually moving from an active-defense strategy of combating adversaries once they reach US networks to a "defend forward" model that operates outside of the United States to deter threats before they reach the United States.[1] Most recently, in a 2023 summary of its cyber strategy, the Defense Department stated that it "will continue to defend forward by disrupting the activities of malicious cyber actors and degrading their supporting ecosystems."[2]

"Defend forward" was not new to the 2023 cyber strategy. For more than five years, the Department has articulated such an operational concept as a key component of its strategy to persistently engage with cyber adversaries.[3] Perhaps most noteworthy about the 2023 strategy summary was the Department's focus on defending forward "in close coordination" with "our global Allies and partners."[4] As the Department observed in its strategy, since 2018, it had "regularly worked with our Allies and partners to help identify vulnerabilities on their government-operated networks," and those activities "have aided U.S. cybersecurity preparedness, contributed to the warfighting capability of the Joint Force, and established or enhanced strong information-sharing relationships with a number of nations, including Ukraine."[5] The Department's term for these operations is "Hunt Forward."[6] As of September 2023, Cyber Command deployed Hunt Forward teams to operations on seventy-seven

1 *See* Dave Weinstein, *The Pentagon's New Cyber Strategy: Defend Forward*, Lawfare (Sept. 21, 2018) ("Whereas active cyber defense, according to the Defense Department's 2011 Strategy for Operating in Cyberspace, consisted of intrusion prevention at the perimeter and 'defeat[ing] adversary activities on DoD networks and systems,' defend forward implies the conduct of operations on non-U.S. networks to 'stop threats before they reach their targets.'"); Jeff Kosseff, *The Contours of "Defend Forward" Under International Law*, Proceedings of the 11th International Conference on Cyber Conflict 13 (2019) ("To be sure, Defend Forward is subject to several legal limits, particularly when it comes to positioning and degradation; but even within these limits, the United States can conduct cyber operations that are far more active than the U.S. active defense concept of years past.").

2 U.S. Defense Department, Summary, 2023 Cyber Strategy of the Department of Defense 6.

3 *See* U.S. Cyber Command, Achieve and Maintain Cyberspace Superiority, Command Vision for U.S. Cyber Command 6 (2018) ("Defending forward as close as possible to the origin of adversary activity extends our reach to expose adversaries' weaknesses, learn their intentions and capabilities, and counter attacks close to their origins.").

4 U.S. Defense Department, *supra* note 2, at 6.

5 *Id*. at 12.

6 *Id*.

networks in twenty-four countries, and Gen. Paul Nakasone, then-commander of Cyber Command, in 2023 called Hunt Forward a "resounding success."[7] For instance, Ukraine credits a Hunt Forward operation conducted before the Russian invasion with helping it maintain train service during the early days of the invasion.[8] And in 2023 alone, Hunt Forward resulted in the release of 90 samples of malware to the public.[9] But what, precisely, is "Hunt Forward"?

Cyber Command defines Hunt Forward operations as "strictly defensive cyber operations conducted by US Cyber Command (USCYBERCOM) at the request of partner nations."[10] If Cyber Command teams are invited by a partner nation, they deploy "to observe and detect malicious cyber activity on host nation networks."[11] Cyber Command reports that Hunt Forward operations "generate insights that bolster homeland defense and increase the resiliency of shared networks from cyber threats."[12] The United States makes Hunt Forward findings public, allowing companies to patch software and "eliminate adversary network accesses and capabilities."[13]

While Cyber Command's general description provides some clues as to the international law issues that might surround Hunt Forward, the inherently sensitive nature of collaborative military cyber operations means that many details cannot be made public. Still, Cyber Command has publicly described many Hunt Forward operations to news outlets and in written statements. Section 2 of this paper reviews those public statements in an attempt to paint a clearer picture of the scope of Hunt Forward. Section 3 applies international law principles to those facts and argues that broader acceptance of collective countermeasures could build on the success of Hunt Forward and similar collaborative cyber operations, allowing more effective responses to internationally wrongful acts of adversaries.

7 Patty Nieberg, *"Hunt Forward" Cyber Teams Have Deployed to 24 Countries, Including Ukraine*, Task and Purpose (Sept. 28, 2023).

8 *Remarks by Assistant Secretary of Defense for Space Policy John Plumb at Center for a New American Security 2023 DOD Cyber Strategy Event*, U.S. Department of Defense (Sept. 12, 2023), https://www.defense.gov/News/Speeches/Speech/Article/3525636/remarks-by-assistant-secretary-of-defense-for-space-policy-john-plumb-at-center/.

9 Posture Statement of General Timothy D. Haugh, Commander, United States Cyber Command, Before the Senate Committee on Armed Services (April 10, 2024) at 7.

10 *Cyber 101: Hunt Forward Operations*, U.S. Cyber Command, https://www.cybercom.mil/Media/News/Article/3218642/cyber-101-hunt-forward-operations/.

11 *Id*.

12 *Id*.

13 *Cyber 101—Defend Forward and Persistent Engagement*, U.S. Cyber Command, (Oct. 25, 2022), https://www.cybercom.mil/Media/News/Article/3198878/cyber-101-defend-forward-and-persistent-engagement/.

2. THE ELEMENTS OF HUNT FORWARD

Cyber Command's public descriptions of Hunt Forward operations help to fill in some of the ambiguities in the general definition of the operations.

A. Searching for Threats and Malware

One of the most common elements in descriptions of Hunt Forward operations is the monitoring of allies' systems and networks for malicious activities. For instance, in its description of a 2020 Hunt Forward operation on the Estonian Defence Force's networks, Cyber Command stated that US and Estonian cyber specialists "hunted for malicious cyber actors on critical networks and platforms."[14] The partner nation can determine the direction of this assessment. In a 2023 Hunt Forward operation in Lithuania, US personnel "analyzed key networks, identified and prioritized by the partner, for evidence of malicious cyber activity while identifying vulnerabilities."[15] Likewise, after a series of Iranian cyber attacks on the Albanian government in 2022, Cyber Command deployed a Hunt Forward team to Albania for three months, "hunting for malicious cyber activity and identifying vulnerabilities on networks of Albania's choice."[16]

In a 2023 interview, Army Maj. Gen. William Hartman, leader of the Cyber National Mission Force, emphasized that this assessment takes place at the invitation of partner nations. And the first step is to detect "anomalous activity," Hartman said. "The team goes through the investigation and at the end of the day, they're going to decide whether there's a potentially malicious IP, or whether the malware that they found, they [want to] know if it's good or bad."[17] Cyber Command is uniquely positioned to provide informed assistance, he said, as it is housed in the same headquarters as the National Security Agency and its Cybersecurity Directorate. "We get access to, to information that the cybersecurity director has, about adversaries that target the United States or allies and partners," Hartman said. "And so ultimately we want to execute an intelligence-driven mission. Because we have intel that tells us that an adversary that threatens us is also threatening one of these partners."[18] This expertise

14 *Hunt Forward Estonia: Estonia, U.S. Strengthen Partnership in Cyber Domain with Joint Operation*, U.S. Cyber Command (Dec. 3, 2020), https://www.cybercom.mil/Media/News/Article/2433245/hunt-forward-estonia-estonia-us-strengthen-partnership-in-cyber-domain-with-joi/#:~:text=3%2C%202020-,Hunt%20Forward%20Estonia%3A%20Estonia%2C%20US%20strengthen%20partnership%20in,cyber%20domain%20with%20joint%20operation&text=Estonian%20and%20U.S.%20cyber%20commands,September%2023%20to%20November%206.

15 *"Building Resilience": U.S. Returns from Second Defensive Hunt Operation in Lithuania*, U.S. Cyber Command (Sept. 12, 2023), https://www.cybercom.mil/Media/News/Article/3522801/building-resilience-us-returns-from-second-defensive-hunt-operation-in-lithuania/.

16 *"Committed Partners in Cyberspace": Following Cyberattack, US Conducts First Defensive Hunt Operation in Albania*, U.S. Cyber Command (Mar. 23, 2023), https://www.cybercom.mil/Media/News/Article/3337717/committed-partners-in-cyberspace-following-cyberattack-us-conducts-first-defens/.

17 Dina Temple-Raston, *Q&A with Gen. Hartman: "There Are Always Hunt Forward Teams Deployed*," The Record (June 20, 2023).

18 *Id*.

uniquely positions Hunt Forward operations to help partners detect threats on their systems and networks.

B. Gathering Intelligence

A primary benefit of Hunt Forward for the United States is to gather intelligence about the cyber tactics of common adversaries and to use that intelligence to improve US cybersecurity. "We do these defend-forward missions, and the whole point of the defend-forward mission is to learn something on someone else's network, a partner network, another nation's network so we can bring back that information and make sure our networks are more secure," Brig. Gen. Reid Novotny, special assistant to the director of Air National Guard for Cybercom, J5, said at a June 2023 conference.[19]

For instance, a Hunt Forward Operation that was conducted after the attack on SolarWinds "yielded eight files attributed to the Russian Intelligence Service (SVR) APT 29" and "yielded information about adversary tactics, techniques, procedures, and intentions," Cyber Command stated.[20] And in its discussion of a joint Hunt Forward operation conducted with the Canadian Armed Forces in Latvia, Cyber Command noted that the operations "provide us advanced notice of adversary tools and techniques."[21]

The intelligence-gathering benefits not only allies but also the United States itself. For instance, in a 2020 *Foreign Affairs* article co-authored with his senior advisor, Michael Sulmeyer, Nakasone wrote that Hunt Forward operations were partly responsible for the United States disrupting "a concerted effort to undermine the midterm elections" in 2018.[22] Likewise, in an April 2022 Senate hearing, Nakasone touted the intelligence value of Hunt Forward as "understanding what our adversaries are doing, and … sharing that broadly, not only with our partners and NATO but also with the private sector."[23] By operating on the systems and networks of partner nations, the United States obtains valuable insights into the methods and techniques of adversaries, and the strategies that the United States develops with partner nations to combat these threats can be useful if the United States later faces similar threats from adversaries.

19 Mark Pomerleau, *US Cyber Command Conducts "Hunt Forward" Mission in Latin America for First Time, Official Says*, DefenseScoop (June 8, 2023).

20 U.S. Cyber Command Public Affairs, *Cyber 101: Hunt Forward Operations*, 960th Cyberspace Wing (Nov. 15, 2022), https://www.960cyber.afrc.af.mil/News/Article-Display/Article/3219164/cyber-101-hunt-forward-operations/.

21 Cyber National Mission Force Public Affairs, *"Shared Threats, Shared Understanding": U.S., Canada and Latvia Conclude Defensive Hunt Operations*, Sixteenth Air Force (Air Forces Cyber) (May 10, 2023), https://www.16af.af.mil/Newsroom/Article/3392740/shared-threats-shared-understanding-us-canada-and-latvia-conclude-defensive-hun/.

22 Paul M. Nakasone & Michael Sulmeyer, *How to Compete in Cyberspace, Cyber Command's New Approach*, Foreign Affairs (Aug. 25, 2020).

23 Transcript of U.S. Senate Committee on Armed Services, Hearing to Receive Testimony on the Posture of United States Special Operations Command and United States Cyber Command in Review of the Defense Authorization Request for Fiscal Year 2023 and the Future Years Defense Program (Apr. 5, 2022) 52.

C. Assisting Allies in Remediation

A key benefit of Hunt Forward for US allies is assistance in remediating harm caused by adversaries. In his 2023 interview, Gen. Hartman of Cyber Command said that Hunt Forward operations involve the use of unclassified equipment on allies' networks. "And when we identify either malware or some type of misconfiguration on a network, we instruct the partner and the partner will take the remediation actions on their own network," he said.[24] He characterized some US remediation support as recommendations made to allies based on best practices.[25]

But what assistance, if any, does the United States provide beyond recommendations for remediation? The public descriptions of Hunt Forward do not provide much more detail. Although Cyber Command describes the operations as "strictly defensive," it is unclear exactly where the line is drawn between defensive and other operations.[26] For instance, in its description of Hunt Forward operations with Ukraine from December 2021 to March 2022, Cyber Command wrote that the United States "conducted network defense activities aligned to critical networks."[27] A 2021 article on Hunt Forward captured the ambiguities in the "strictly defensive" terminology: "The fact is it can be difficult to draw a hard line between offense and defense in cyberspace," Brad D. Williams wrote on the news website Breaking Defense. "For instance, if CYBERCOM disrupts an adversary's infrastructure ahead of a suspected attack against the US, is that an offensive or a defensive operation?"[28] Williams reported that Air Force Lt. Gen. Charles Moore, the Cyber Command deputy commander, "likened CYBERCOM's evolution from that of a football team where only the offense or defense is on the field at one time to more like a hockey team, where any given line change plays both an offensive and defensive role."[29]

In short, there is no evidence in the public record that US remediation assistance goes beyond providing technical recommendations and assistance for partner nations to harden their defenses against adversaries. But any legal analysis of the *potential* of Hunt Forward and future collaborative efforts should consider possible impacts of the operations on adversaries' systems. One component of Defend Forward is "positioning," which Cyber Command describes as "a forward cyber posture that can

24 Temple-Raston, *supra* note 17.

25 *Id*.

26 Hunt Forward Operations are generally characterized as fitting within the Defensive Cyberspace Operations-Internal Defensive Measures mission. *See* Paul Schuh, *Expeditionary Cyberspace Operations*, The Cyber Defense Review at 37 (Spring 2023). If an operation is not on "friendly cyber-space terrain," but instead "is conducted external to the defended network, in foreign cyberspace, and without the permission of the affected system's owner," it falls within the Defensive Cyberspace Operations-Response Action mission. Air Force Doctrine Publication 3-12, Cyberspace Operations at 8 (2023). External effects operations also can fall within the Offensive Cyberspace Operations mission.

27 Cyber National Mission Force Public Affairs, *Before the Invasion: Hunt Forward Operations in Ukraine*, U.S. Cyber Command (Nov. 28, 2022), https://nsarchive.gwu.edu/sites/default/files/documents/rmsj3h-751x3/2022-11-28-CNMF-Before-the-Invasion-Hunt-Forward-Operations-in-Ukraine.pdf.

28 Brad D. Williams, *CYBERCOM Has Conducted "Hunt-Forward" Ops in 14 Countries, Deputy Says*, Breaking Defense (Nov. 10, 2021).

29 *Id*.

be leveraged to persistently degrade the effectiveness of adversary capabilities and blunt their actions and operations before they reach U.S. networks."[30] Such activities should be part of future collaboration, allowing the United States and its allies to blunt the impact of persistent attacks by common adversaries.

3. HUNT FORWARD, INTERNATIONAL LAW, AND COLLECTIVE COUNTERMEASURES

This section analyzes the permissibility of Hunt Forward operations under international law and explores the potential for more robust collaboration between the United States and its allies in a fight against common adversaries. As seen above, nothing in the public record clearly defines the boundaries of Hunt Forward operations. Nor does anything suggest that Hunt Forward operations have a direct impact outside of the partner countries.

Part A of this section examines the international legal issues surrounding of the conduct of Hunt Forward operations on partner nations' physical territory, systems, and networks, and the importance of proper authorization. Part B examines the trickier international legal issues surrounding any impacts of collaborative operations on adversaries and argues that greater acceptance of collective countermeasures would give the United States and its allies more flexibility to take full operational advantage of Hunt Forward in the event of an internationally wrongful act by an adversary.

A. International Legal Issues Surrounding Partner Nations

Under Hunt Forward, US forces conduct cyber operations within the physical territory of allies and might monitor their systems or networks to identify and remediate adversarial threats. These actions are permissible under international law because the United States operates within the parameters of the consent that the partner nation provides.

Assessing Hunt Forward requires an examination of whether the operations breach any international legal obligations that the United States owes to the partner nations.[31] Merely analyzing allies' systems and networks with their consent does not raise concerns under international law. But issues might arise if US Hunt Forward operations inadvertently cause damage within the systems or networks of the partner nation.[32] Here it is vital that the United States receive express authorization from the

30 *See* Kosseff, *supra* note 1, at 4.

31 *Id*.

32 Government Offices of Sweden, Position Paper on the Application of International Law in Cyberspace 2 (July 2022) ("In general, Sweden is of the view that violations of sovereignty may arise from cyber operations that result in damage or loss of functionality. Altering and interfering with data without causing physical harm may also violate sovereignty.").

partner to conduct cyber operations within its territory.[33] To be sure, at least some states hold that some minor cyber harms, even without consent, do not automatically lead to sovereignty violations.[34]

Even with the potential allowance of cyber operations that cause insignificant harm, Hunt Forward operations must always be based on clearly defined consent and not exceed the authorized scope of that consent.[35] The United States should take great care to ensure that authorization is clearly defined and describes each specific part of the systems, networks, and information that US personnel may access. The authorization should also specify the types of activities that are permissible under Hunt Forward and the aims and duration of the operation.

B. International Legal Issues Surrounding Adversaries and the Case for Collective Countermeasures

Although Hunt Forward operations are conducted from the physical territory of partner nations and are characterized as strictly defensive, any legal analysis of the potential of future collaborations should account for possible impacts on the adversaries' network.

Merely remediating and preventing further harm to the systems of partner nations might frustrate the objectives of adversaries, but it is difficult to see how that assistance would violate international legal obligations owed to adversaries. Although Hunt Forward might disturb adversaries' objectives in cyberspace, helping their targets harden their defenses does not raise any reasonable concerns under international law. Any legal analysis from the perspective of adversaries should focus on impacts on the adversaries' systems, networks, and information.

33 *See* Italian Position Paper on "International Law and Cyberspace" 4 ("Italy finds that, according to the same principle, a State may not conduct cyber operations from the territory of another State without its express authorization.").

34 *See* Government of Canada, International Law Applicable in Cyberspace 17 ("The rule of territorial sovereignty does not require consent for every cyber activity that has effects, including some loss of functionality, in another State. Activities causing negligible or de minimis effects would not constitute a violation of territorial sovereignty regardless of whether they are conducted in the cyber or non-cyber context."); Finnish Government, Finland Published Its Positions on Public International Law in Cyberspace (Oct. 15, 2020) ("The assessment of whether an unauthorized cyber intrusion violates the target State's sovereignty depends on the nature and consequences of the intrusion. The matter is subject to a case-by-case assessment."). To be sure, "there is no clear consensus as to whether an act of cyber aggression could constitute a standalone violation of sovereignty, or if it must implicate another rule such as non-intervention." Jeff Kosseff, *Retorsion as a Response to Ongoing Malign Cyber Operation*, Proceedings of the 12th International Conference on Cyber Conflict (2020) at 12.

35 *See* Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations 27 (2017) [hereinafter Tallinn Manual]. ("Consider a case in which non-State actors are engaged in harmful cyber activities on a State's territory against that State. The State in question does not have the technical ability to put an end to those activities and therefore requests the assistance of another State. The assisting State's ensuing cyber operations on the other State's territory would not violate the latter's sovereignty as long as the operations remain within the scope of its consent."); Harriet Moynihan, The Application of International Law to State Cyberattacks, Sovereignty and Non-intervention 48 (Chatham House, Dec. 2019) ("A violation of sovereignty occurs when one state exercises authority in another state's territory without consent in relation to an area over which the territorial state has the exclusive right to exercise its state powers independently.").

The public descriptions of Hunt Forward operations suggest that one of the key benefits of Hunt Forward to the United States is learning about adversaries' cyber tactics. Such information-gathering—occurring on the physical territory of partner nations—does not violate adversaries' sovereignty. Under the majority view of international law, the United States would not violate sovereignty even if the Hunt Forward operations resulted in the United States observing the adversary's systems.[36] Although cyber espionage is not per se a violation of international law, the operations could cross the line to a sovereignty violation if they cause sufficient damage.[37] To be clear, the descriptions of Hunt Forward operations do not suggest espionage activities within the systems of adversaries, much less operations that cause damage to those systems. But such considerations are important when examining the permissible legal scope of future mutual cyber operations.

Likewise, the descriptions of Hunt Forward do not suggest that US Hunt Forward operations include helping partner nations to penetrate adversaries' systems and disable the capabilities that are at the source of the malign cyber acts. But such actions would be a logical extension of Hunt Forward in future collaborative operations and allow both the United States and its allies to work together to cause adversaries to cease persistent cyber aggression that results in internationally wrongful acts. Such operations might raise concerns about infringements of the adversaries' sovereignty.[38] Accordingly, collaboration between the United States and its allies must either continue to avoid such operations on adversaries' networks or be grounded in a legal justification that permits activities that would otherwise violate an international legal obligation owed to the adversary.

A plausible legal justification for such activities would be countermeasures, which, according to Michael Schmitt and Sean Watts, are "non-forcible, but otherwise

36 *See* Tallinn Manual, *supra* note 35, at 168 ("Although peacetime cyber espionage by States does not per se violate international law, the method by which it is carried out might do so."); New Zealand, The Application of International Law to State Activity in Cyberspace ¶ 14 (2020) ("There is a range of circumstances—in addition to pure espionage activity—in which an unauthorized cyber intrusion, including one causing effects on the territory of another state, would not be internationally wrongful."); Government of Canada, *supra* note 34, at ¶ 19 ("Importantly, some cyber activities, such as cyber espionage, do not amount to a breach of territorial sovereignty, and hence to a violation of international law. They may, however, be prohibited under the national laws of a State."); but see Costa Rica's Position on the Application of International Law in Cyberspace ¶ 22 ("Furthermore, surveillance operations may be carried out in ways that lead to breaches of State sovereignty or other rules of international law. As such, Costa Rica believes that, in some circumstances, cyber espionage may amount to a breach of State sovereignty.").

37 *See* Tallinn Manual, *supra* note 35, at 170 ("For instance, if organs of one State, in order to extract data, hack into the cyber infrastructure located in another State in a manner that results in a loss of functionality, the cyber espionage operation violates, in the view of the Experts, the sovereignty of the latter.").

38 *See id*. at 21 ("The Experts agreed that, in addition to physical damage, the remote causation of loss of functionality of cyber infrastructure located in another State sometimes constitutes a violation of sovereignty, although no consensus could be achieved as to the precise threshold at which this is due to the lack of expressions of *opinio juris* in this regard."); Germany, On the Application of International Law in Cyberspace, Position Paper 4 (2021) ("If functional impairments result in substantive secondary or indirect physical effects in the territory of the target State (and a sufficient causal link to the cyber operation can be established), a violation of territorial sovereignty will appear highly probable.").

unlawful, acts undertaken in response to another state's breach of an international law obligation."[39] The commentary to the 2001 Draft Articles on Responsibility of States for Internationally Wrongful Acts states that "the commission by one State of an internationally wrongful act may justify another State injured by that act in taking non-forcible countermeasures in order to procure its cessation and to achieve reparation for the injury."[40]

When states exercise countermeasures, they face important limitations. A state that is exercising countermeasures "may only take countermeasures against a State which is responsible for an internationally wrongful act in order to induce that State to comply with its obligations" under international law.[41] Among the many requirements of countermeasures is that they "are limited to the non-performance for the time being of international obligations of the State taking the measures towards the responsible State" and that they must "be taken in such a way as to permit the resumption of performance of the obligations in question."[42] Countermeasures are also subject to the rule of proportionality, meaning that they "must be commensurate with the injury suffered, taking into account the gravity of the internationally wrongful act and the rights in question."[43]

Accordingly, for the purposes of the discussion of countermeasures in this paper, let us assume that the adversary has committed an internationally wrongful act against the US ally. For instance, imagine that State A maintains an ongoing denial-of-service attack against government servers in State B, a US ally. Assuming that the denial-of-service attack constitutes a breach of international legal obligations, State B would be entitled to engage in proportional countermeasures against State A, with the goal of terminating the internationally wrongful acts.

Indeed, states widely recognize the availability of countermeasures to respond to internationally wrongful acts in cyberspace,[44] and the *Tallinn Manual* takes a similar stance.[45] While the ability of an injured state to exercise cyber countermeasures is generally accepted, a more disputed issue is whether another state can lawfully exercise cyber countermeasures on behalf of the injured state. In other words, could the United

39 Michael N. Schmitt & Sean Watts, *Collective Cyber Countermeasures?* 12 Harvard Nat. Sec. J. 373, 377 (2021). To the extent that the activities were unfriendly but did not violate an international legal obligation, they could be justified as retorsion (*see* Tallinn Manual, *supra* note 35, at 112), but the sovereignty issues surrounding damage to an adversary's computer make countermeasures a more likely justification.

40 Int'l Law Comm'n, Draft Articles on Responsibility of States for Internationally Wrongful Acts, Rep. of the Int'l Law Comm'n on the Work of Its Fifty-Third Session, U.N. Doc. A/56/10, at 75 (2001) [hereinafter Draft Articles].

41 Draft Articles, note 40, at 129.

42 *Id*.

43 *Id*. at 134.

44 *See, e.g.*, New Zealand, The Application of International Law to State Activity in Cyberspace ¶ 21 (2020 (Countermeasures "may include, but are not limited to, cyber activities that would otherwise be prohibited by international law.").

45 Tallinn Manual, *supra* note 35, at 111 ("A State may be entitled to take countermeasures, whether cyber in nature or not, in response to a breach of an international legal obligation that is owed by another State.").

States engage in countermeasures against State A on behalf of its injured ally, State B? The Draft Articles touch on issues related to such "collective countermeasures" but do not take an explicit position on their permissibility.[46] Article 48 allows a non-injured state "to invoke the responsibility of another State" if "the obligation breached is owed to the international community as a whole," but that does not explicitly address collective countermeasures.[47] In the Draft Articles commentary, the International Law Commission asserted that "there appears to be no clearly recognized entitlement of States referred to in article 48 to take countermeasures in the collective interest" and that, therefore, "it is not appropriate to include in the present articles a provision concerning the question whether other States, identified in article 48, are permitted to take countermeasures in order to induce a responsible State to comply with its obligation."[48] James Crawford, the International Law Commission's special rapporteur at the time the articles were drafted, later wrote that a proposal for collective countermeasures was too divisive for inclusion.[49]

Most national position statements about international law in cyberspace say nothing about collective countermeasures. The experts who drafted the *Tallinn Manual* were divided as to their permissibility. Most of the experts concluded that "purported countermeasures taken on behalf of another State are unlawful," but a minority concluded that "a non-injured State may conduct countermeasures as a response to an internationally wrongful act committed against an injured State so long as the latter request that it do so."[50] The experts were more closely divided as to whether "a State may assist another State in conducting the latter's countermeasures."[51]

The question of the permissibility of collective countermeasures reemerged nearly twenty years after the publication of the Draft Articles and two years after the publication of the second edition of the *Tallinn Manual*. At the 2019 International Conference on Cyber Conflict (CyCon), Estonian president Kersti Kaljulaid announced Estonia's stance that "states which are not directly injured may apply countermeasures to support the state directly affected by the malicious cyber operation."[52] Since then, some other countries have embraced that position. In its December 2020 statement on international law in cyberspace, New Zealand said that it was "open to the

46 *See* Jeff Kosseff, *Collective Countermeasures in Cyberspace*, 10 Notre Dame J. Int'l & Comp. Law 18, 24 (2020) ("The lengthy and spirited debate is evident in the text of the Draft Articles, which do not directly address the legality of collective countermeasures, but dance around the issue quite a bit.").

47 Draft Articles, *supra* note 40, at 126.

48 *Id*. at 139.

49 James Crawford, *The ILC's Articles on Responsibility of States for Internationally Wrongful Acts: A Retrospect*, 96 Am. J. Int'l L. 874, 884 (2002). ("Although the proposal received a degree of support both within and outside the ILC, some governments strongly opposed it. In the end, discretion seemed the better part of valor, particularly having regard to the interaction of these issues with the general mandate of the Security Council.").

50 Tallinn Manual, *supra* note 35, at 132.

51 *Id*.

52 *President Kaljulaid at CyCon 2019: Cyber Attacks Should Not Be an Easy Weapon*, ERR News (May 29, 2019), https://news.err.ee/946827/president-kaljulaid-at-cycon-2019-cyber-attacks-should-not-be-easy-weapon.

proposition that victim states, in limited circumstances, may request assistance from other states in applying proportionate countermeasures to induce compliance by the state acting in breach of international law."[53] In a stronger endorsement of collective countermeasures, Costa Rica maintained in its cyber law position statement that "States may respond collectively to cyber or non-cyber operations that amount to internationally wrongful acts, by resorting to cyber or non-cyber countermeasures."[54] And Ireland stated in 2023 that collective countermeasures "are permissible in limited circumstances."[55]

Some countries have questioned or rejected Estonia's position on collective countermeasures. Canada, while open to the general concept of assisting an injured state, noted that it considered collective countermeasures but "does not, to date, see sufficient State practice or *opinio juris* to conclude that these are permitted under international law."[56] And France went further in its refusal to recognize the concept, stating that "collective counter-measures are not authorised, which rules out the possibility of France taking such measures in response to an infringement of another State's rights."[57]

The success of Hunt Forward weighs in favor of broader global acceptance of collective cyber countermeasures. To be sure, the United States has consistently characterized Hunt Forward as purely defensive and has not described any operations that would need to be justified as countermeasures. But acceptance of collective countermeasures in cyberspace would provide such operations with breathing space to collaborate more effectively.[58] Collective countermeasures would allow collaborative operations to expand from merely helping partners identify and analyze threats on their systems, such as Hunt Forward, to also helping the partners stop malicious activities at their source.

For instance, consider a small state whose local-government computer systems are routinely targeted by malicious code transmitted by a larger adversarial nation. The malware often prevents the local governments from conducting their daily business and serving constituents. Such malign actions likely violate international legal obligations and would entitle the target state to engage in limited and proportionate

53 New Zealand, The Application of International Law to State Activity in Cyberspace ¶ 22 (2020).

54 Costa Rica's Position, *supra* note 36, at ¶ 15.

55 Ireland Position Paper on the Application of International Law in Cyberspace ¶ 26 ("The possibility of imposing third party or collective countermeasures in the cyber context is particularly relevant for states that may consider it necessary to respond to a malicious cyber-operation with a counter-operation, but lack the technological capacity to do so on their own.").

56 Government of Canada, *supra* note 34, at ¶ 37. Canada took a middle ground, reasoning that "assistance can be provided on request of an injured State, for example where the injured State does not possess all the technical or legal expertise to respond to internationally wrongful cyber acts. However, decisions as to possible responses remain solely with the injured State." *Id*.

57 France, International Law Applied to Operations in Cyberspace, Paper shared by France with the Open-Ended Working Group Established by Resolution 75/240 at 4.

58 *See* Schmitt & Watts, *supra* note 39, at 410 ("The unique nature of cyberspace suggests a need for greater tolerance of countermeasures.").

countermeasures intended to terminate the malign actions. For example, the small state might remotely disable the adversary's computer systems that are the source of the malware. But imagine that the small state lacks the skills, knowledge, and staffing to implement such an operation.[59] Under the doctrine of collective countermeasures, US Hunt Forward teams could either directly conduct the operation against the adversary or assist the small state in doing so. Such an operation not only would benefit the small state by stopping the malign operations on its systems but also would benefit the United States by weakening the source of a potential future operation against US systems.

A legitimate criticism of collective cyber countermeasures is that they are susceptible to abuse and could escalate tensions. While such concerns are understandable, they could be mitigated by the fact that the same limits that are imposed on countermeasures in the offline world would apply in cyberspace. For instance, countermeasures must be "proportional," meaning "commensurate with the injury suffered, taking into account the gravity of the internationally wrongful act and the rights in question."[60] In the example above, a permissible countermeasure might include knocking an adversary's computer offline if that computer had been the source of the malware, but it would not be proportional to mount a broader attack on a larger telecommunications system. The collective countermeasures can only have the purpose of causing the adversary to "comply with its obligations" under international law,[61] and the states must terminate their countermeasures "as soon as the responsible State has complied with its obligations" under international law.[62] In other words, collective cyber countermeasures would not be a blank check for non-injured states to attack adversaries and escalate tensions.[63]

To be sure, my proposal would require a significant expansion of collaboration beyond the current, purely defensive Hunt Forward construct. It would require different personnel, moving beyond only the Cyberspace Protection Teams that focus on defending cyberspace and toward teams that work on Defensive Cyberspace Operations-Response Actions or Offensive Cyberspace Operations. While the legal issues surrounding my proposal are more complex and the risk of escalation increases, the success of the current Hunt Forward model suggests that the United States and its allies have good reason to embrace the model of collective countermeasures and collaborate with allies not only in gathering information and fixing harm but by preventing further aggression by adversaries.

59 *Id*. at 377–78 ("The lack of collective responses to international law breaches would render self-help through countermeasures impossible for many weak states. If forced to respond alone, they would not be able to induce more powerful responsible states to cease unlawful activity.").

60 Draft Articles, *supra* note 40, at 134.

61 *Id*. at 129.

62 *Id*. at 137.

63 *See* Kosseff, *supra* note 46, at 32 ("That is why collective countermeasures would be subject to all of the limitations that apply to countermeasures taken by the target state. It also would be reasonable to impose additional responsible limits on third parties seeking to engage in collective countermeasures.").

4. CONCLUSION

The first five years of Hunt Forward operations have demonstrated substantial benefits not only for partner nations but also for the United States. By helping allies identify the source of malicious cyber operations on their networks, the United States gains valuable intelligence that it can use on domestic security. Provided that the United States has clear and specific authorization from the partner nation, Hunt Forward operations, as publicly described, do not raise concerns under international law. Broader acceptance of collective countermeasures would enable the United States and its partners to further leverage collaboration to degrade the capability of adversaries. While concerns about the misuse of collective countermeasures are legitimate, the international community could address many of those concerns by applying the same limits that nations face under the general law of countermeasures, including proportionality and limits on purpose and duration. Expanding collaboration beyond Hunt Forward, through the embrace of collective countermeasures, would more fully realize the benefits of Defend Forward and persistent engagement.

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

Specially Affected States' Push for Collective Countermeasures

Lisandra Novo
Staff Lawyer
Strategic Litigation Project, Atlantic Council
Washington, DC, United States

Abstract: At CyCon 2019, Estonia publicly affirmed its position on the permissibility of collective countermeasures as a response to malicious cyber operations. While Estonia was the first State to publicly express its position on the topic, several other States have now also done so in the five years since. Some, like Ireland and Costa Rica, support the position that States may engage in collective countermeasures under certain circumstances. Others, like Denmark and the United Kingdom, believe that the question remains unsettled. At the other end of the spectrum, France has adopted the position that collective countermeasures are prohibited under international law.

As more than twenty States have publicly adopted a position on the permissibility of countermeasures in response to malicious cyber operations, we seem to be in the nascent stage of an emerging norm in international law applicable to cyberspace. In this paper, I summarize publicly available State positions on collective countermeasures to show that the question of their legality is at least an open one. I also call for specially affected States to be given due consideration in the formation of custom around this issue. I attempt to trace the origins of the French position and argue that it is based on an exceedingly narrow and outdated interpretation of international law. Lastly, I argue that Estonia's position promotes peace and security by allowing States that may not have the technological capability to individually respond to malicious cyber operations with countermeasures to seek assistance rather than having to resort to force.

Keywords: *collective countermeasures, specially affected States, Estonia, France, malicious cyber operations, law of State responsibility*

1. INTRODUCTION

Costa Rica is one of the latest States to adopt the view that non-injured States may engage in collective countermeasures to assist a State that has been the victim of malicious cyber operations attributable to another State. Notably, it has taken this position as a victim State that suffered a massive ransomware attack and had to seek assistance from other States. It joins Estonia, the first State to publicly espouse a position on the permissibility of collective countermeasures five years ago. Estonia is another small State and was the victim of what is considered the first cyberattack against a State. Since then, more than twenty States have expressed their views on the law applicable to countermeasures in response to malicious cyber operations.

In this paper, I outline the various available State positions on countermeasures, from those that remain silent on the issue of collective responses and those that frame it as an open question to those that take a firm stance for or against. The growing number of States advocating for collective countermeasures, particularly those that can be classified as specially affected States in the formation of custom, points to an emerging norm of international law. I then examine the law of State responsibility to demonstrate that the position that collective countermeasures are not allowed is unfounded and that, at a minimum, the question is not settled. Lastly, I argue that the position advanced by States like Estonia and Costa Rica—to allow States to engage in collective countermeasures—promotes peace and security. It does so, I maintain, by giving victim States that do not independently have the resources to adequately respond to cyberattacks an effective and practical solution without having to resort to force to defend their interests.

2. EMERGING NORM ON PERMISSIBLE RESPONSES TO CYBERATTACKS

The increasing number of States supporting collective countermeasures as a permissible response to cyberattacks points to a new norm emerging in international law. Moreover, many of these States are specially affected States in the formation of custom that should be given additional consideration as the norm develops. It is in the best interest of other specially affected States, like small States or States in the global majority that do not independently have the technical capacity to carry out countermeasures, to follow Estonia and Costa Rica's example and express their positions on this question.

A. Public Positions

As of March 2024, twenty-five States have addressed the issue of the legality of countermeasures as a response to cyberattacks, ten of which have explicitly addressed the question of collective countermeasures.[1] Of those ten States, six—Costa Rica, Estonia, Ireland, New Zealand, Poland, and Romania—have said that, at least in some circumstances, international law as it stands today allows more than one State, including non-injured States, to carry out countermeasures in response to malicious cyber activities.[2] Two States—Denmark and the United Kingdom—have said that the question remains open and requires further consideration.[3] Canada has taken the nuanced position that it "does not, to date, see sufficient State practice or *opinio juris* to conclude that these are permitted under international law."[4] Only France has gone so far as to say that collective countermeasures are "not authorized" under international law.[5] Brazil appears to be the only State that has questioned the customary status of countermeasures altogether, even for the injured State.[6]

Estonia was the first State to publicly advocate for collective countermeasures.[7] In her speech at CyCon 2019, Kersti Kaljulaid, the then-president of Estonia, pointed out that States may respond individually to malign cyber activity through diplomatic measures and countermeasures and force in self-defense.[8] She then announced that Estonia was "furthering" the position that countermeasures should be considered permissible collective responses.[9] As other authors have noted, this position needed "furthering" because, while the legality of collective diplomatic responses and collective self-defense is well established, the question of collective countermeasures has not been settled.[10]

1 *See Countermeasures*, Int'l Cyber Law: Interactive Toolkit, https://cyberlaw.ccdcoe.org/w/index.php?title=Countermeasures&oldid=3892 (last visited Dec. 10, 2023) (for States that have expressed public positions on countermeasures but remained silent on collective countermeasures, see positions of Australia, Brazil, Finland, Germany, Israel, Italy, Japan, the Netherlands, Norway, Russia, Singapore, Sweden, Switzerland, and the U.S.). *See also Czechia Has Published a Position Paper on the Application of International Law in Cyberspace*, Ministry of Foreign Affairs of the Czech Republic (Feb. 27, 2024), https://mzv.gov.cz/jnp/en/foreign_relations/international_law/news/czechia_has_published_a_position_paper.html.

2 *See* positions of Estonia, Romania, New Zealand, Poland, Costa Rica, and Ireland in *Countermeasures, supra* note 1.

3 *See* positions of Denmark and the U.K. in *Countermeasures, supra* note 1.

4 *International Law Applicable in Cyberspace*, Gov't Canada (Apr. 22, 2022), https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/peace_security-paix_securite/cyberspace_law-cyberespace_droit.aspx?lang=eng.

5 *See* France's 2019 position in *Countermeasures, supra* note 1; *Droit International Appliqué aux Opérations dans le Cyberespace*, Ministère des Armées (2023), https://www.defense.gouv.fr/comcyber/droit-international-applique-aux-operations-cyberespace (last visited Dec. 10, 2023) [hereinafter *France 2023 Position on International Law Applicable to Cyber Operations*].

6 *See* Brazil's position in *Countermeasures, supra* note 1.

7 Michael Schmitt, *Estonia Speaks Out on Key Rules for Cyberspace*, Just Security (June 10, 2019), https://www.justsecurity.org/64490/estonia-speaks-out-on-key-rules-for-cyberspace/ [hereinafter Schmitt, *Estonia Speaks Out*].

8 NATO CCDCOE, *Keynote Address by H.E. Kersti Kaljulaid, President of the Republic of Estonia—CyCon 2019*, YouTube (Aug. 9, 2019), https://youtu.be/tdWPjEKARVA?feature=shared.

9 *Id*.

10 Schmitt, *Estonia Speaks Out, supra* note 7.

In addition to their individual positions, States are also addressing this issue through intergovernmental organizations. For example, NATO secretary general Jens Stoltenberg acknowledged that NATO "need[s] a full spectrum response" to "serious cyber-attacks even if they don't cross the [armed attack] threshold."[11] In October 2023, at the first-ever International Conference of the European Union in the Legal Sphere of Cyber Defence, EU member States' cyber commanders met to discuss legal issues relating to cyber defense. They concluded that the "promotion of common positions, as well as coordinated countermeasures, should be considered the most powerful tool in establishing a useful framework for deterring malicious actors and being prepared to respond effectively to such threats."[12] Participants also concluded that there "are no legal obstacles to developing collective countermeasures as a way of dealing with malicious activities in cyberspace."[13]

Some scholars' positions have also evolved. Michael Schmitt, general editor of the *Tallinn Manual*, wrote in 2014 that collective countermeasures were not permissible.[14] By 2021, Schmitt had adopted the position that, even though the question of their legal permissibility remained unsettled, accepting collective countermeasures as a lawful response to cyberattacks was the better option.[15] Schmitt also noted that, for a country like Estonia that may not have the individual cyber capability to respond effectively to cyberattacks, collective countermeasures under an alliance like NATO would be a logical solution.[16] The same could be said about regional organizations like the EU or the African Union, which recently released its common position on the application of international law in cyberspace.[17]

B. Specially Affected States

The paradigm that the only permissible collective responses are diplomatic measures or self-defense leaves States on their own if diplomatic measures are not sufficient

11 Jens Stoltenberg, *Stoltenberg Provides Details of NATO's Cyber Policy*, Atlantic Council (May 16, 2018), https://www.atlanticcouncil.org/blogs/natosource/stoltenberg-provides-details-of-nato-s-cyber-policy/.

12 Ministerio de Defensa, *Balance of the "EU International Conference in the Legal Field of Cyber Defense,"* Spain (Oct. 2, 2023), https://emad.defensa.gob.es/en/prensa/noticias/2023/10/Listado/231002-ni-balance-mcce-conferencia-internacional.html.

13 *Id*.

14 Michael N. Schmitt, *"Below the Threshold" Cyber Operations: The Countermeasures Response Option and International Law*, 54 Va. J. Int'l L. 697, 731 (2014).

15 *E.g.*, Michael Schmitt, *Three International Law Rules for Responding Effectively to Hostile Cyber Operations*, Just Security (July 13, 2021), https://www.justsecurity.org/77402/three-international-law-rules-for-responding-effectively-to-hostile-cyber-operations/; Michael N. Schmitt & Sean Watts, *Collective Cyber Countermeasures?*, 12 Harv. Nat'l Sec. J. 373, 380 (2021). *See also* Michael Schmitt, *International Law at NATO's Brussels Summit*, EJIL: Talk! (June 30, 2021), https://www.ejiltalk.org/international-law-at-natos-brussels-summit/ (arguing that NATO allies could use collective countermeasures to respond to subthreshold armed attacks); Franklin D. Kramer, Hans Binnendijk & Lauren M. Speranza, NATO Priorities After the Brussels Summit (Atlantic Council 2018) (arguing NATO should develop a doctrine to respond to malicious operations that includes collective countermeasures).

16 Schmitt, *Estonia Speaks Out, supra* note 7.

17 The common position did not address countermeasures. *See* Mohamed Helal, *Common African Position on the Application of International Law to the Use of Information and Communication Technologies in Cyberspace, and All Associated Communiqués Adopted by the Peace and Security Council of the African Union (Feb. 2, 2024)*, Ohio State Legal Studies Research Paper No. 823, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4714756.

to address a violation of international law that does not constitute an armed attack. Then consider that the majority of cyberattacks do not reach the threshold of an armed attack.[18] States that do not have the independent capacity to respond to such attacks but that have the resources to hire private actors as their agents are free to do so, as the actions of those private actors would be attributable to them.[19] Yet they cannot turn to other States. This means that States that are still building up their cyber capabilities but are unable or unwilling to hire private actors lack effective recourse.

It is not surprising that Estonia opened the debate on collective countermeasures, as it is widely considered to be the first State victim of a politically motivated cyberattack.[20] Another State pushing for collective countermeasures, Costa Rica, was the victim of a 2022 ransomware attack that caused such extensive and long-term damage to government agencies that the country declared a national emergency.[21] Researchers later warned that malicious actors have intentionally begun targeting States in the global majority more frequently.[22] The EU also recently expressed its concern about the threat of ransomware, highlighting the "blurring of the lines between state-sponsored and criminal or financially motivated actors."[23]

As ten States have taken differing positions on the question of the legality of collective countermeasures and there is no treaty or convention on the subject, it is important to assess whether a customary norm is beginning to take shape and, if so, in what direction. To determine whether something constitutes customary international law, it is necessary to look to State practice and *opinio juris*—what a State says it understands as legal obligations.[24] As Schmitt and Vihul note, it is difficult to determine the precise moment a nascent norm relating to cyber activities crystallizes into a customary rule, partly because much of what States do in the cyber realm is not visible to the general public and States are often hesitant to publicly opine on the legality of certain actions.[25] Nevertheless, as Michael Wood, special rapporteur on the identification of customary international law, has explained, for a customary international law rule "to emerge or be identified," "the practice need not be unanimous (universal); but, it

18 *See, e.g.*, Henning Lahmann, Unilateral Remedies to Cyber Operations: Self-Defence, Countermeasures, Necessity, and the Question of Attribution 113–178 (2020).

19 *See, e.g.*, Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations 131 (Michael N. Schmitt ed., 2nd ed., 2017).

20 *Estonian Denial of Service Incident*, Council on Foreign Relations (May 2007), https://www.cfr.org/cyber-operations/estonian-denial-service-incident.

21 Kate Conger & David Bolaños, *Russian Hacking Cartel Attacks Costa Rican Government Agencies*, N.Y. Times (May 17, 2022), https://www.nytimes.com/2022/05/17/us/politics/russia-hacking-costa-rica.html.

22 Janosch Delcker, *Ransomware: Cyber Criminals Are Coming for the Global South*, Deutsche Welle (Aug. 28, 2022), https://www.dw.com/en/ransomware-cyber-criminals-are-coming-for-the-global-south/a-62917234.

23 *EU Statement—UN Open-Ended Working Group on ICT: Existing and Potential Threats*, Eur. External Action Serv. (Mar. 5, 2024), https://www.eeas.europa.eu/delegations/un-new-york/eu-statement-—-un-open-ended-working-group-ict-existing-and-potential-threats-0_en.

24 *Id*. conclusion 2.

25 Michael N. Schmitt & Liis Vihul, *The Nature of International Law Cyber Norms, in* CCDCOE Tallinn Paper No. 5 (2014), 26–28.

must be 'extensive' or, in other words, sufficiently widespread."[26] While there is no specific number of States that need to engage in a practice for a norm to crystallize, participation "must also be broadly representative and include those States whose interests are specially affected."[27] Those are States for which the stakes are higher in the resolution of a particular question.[28]

Consideration of the role of specially affected States in the formation of custom is most commonly associated with the *North Sea Continental Shelf* cases judgment. In that judgment, the International Court of Justice (ICJ) gave special weight to the practice of coastal States over landlocked States on the question of maritime delimitation.[29] There, the Court held that even a short period of time would not necessarily prevent a norm from becoming custom so long as "State practice, including that of States whose interests are specially affected, should have been both extensive and virtually uniform" and that it should "have occurred in such a way as to show a general recognition that a rule of law or legal obligation is involved."[30] In that sense, as Special Rapporteur Wood explained, any assessment must take specially affected States' practice into account, and "such practice should weigh heavily (to the extent that, in appropriate circumstances, it may prevent a rule from emerging)."[31]

Which States count as "specially affected" depends on the rule in question.[32] For example, the International Law Commission (ILC) Draft Conclusions explain that for a rule on foreign investment, the practice of capital-exporting States and States where the investment is made should be considered.[33] The term does not, however, "refer to the relative power of States."[34] Small States or other States that may not have the technological capacity to respond to cyberattacks independently are the kind most likely to request assistance to respond to a subthreshold cyberattack. States that have the capacity to individually engage in countermeasures and that could be called upon for assistance, while analogous to capital-exporting States in the above example, could simply deny these requests if they wished.[35] Small and low-income States without

26 Second Report on Identification of Customary International Law by Michael Wood, Special Rapporteur, ¶ 52, *in* Int'l Law Comm'n, Rep. on the Work of Its Sixty-sixth Session, U.N. Doc. A/CN.4/672 (May 22, 2014) [hereinafter 2014 Special Rapporteur Report on Identification of Customary International Law].

27 *Id. See also* Schmitt & Vihul, *supra* note 25, at 23.

28 Draft Conclusions on Identification of Customary International Law, with Commentaries, conclusion 8, commentary, ¶ 4, *in* Int'l Law Comm'n, Rep. on the Work of Its Seventieth Session, 117 U.N. Doc. A/73/10 (2018) [hereinafter ILC Draft Conclusions]. The term "specially affected State" here should not be confused with its use in the context of the adjudication of State responsibility, which typically refers to an injured State.

29 North Sea Continental Shelf Cases (Ger./Den.; Ger./Neth.), Judgment, 1969 I.C.J. Rep. 3, ¶¶ 70–74 (Feb. 20).

30 *Id*. ¶ 74.

31 2014 Special Rapporteur Report on Identification of Customary International Law, *supra* note 26, ¶ 54.

32 *Id*.

33 ILC Draft Conclusions, *supra* note 28, conclusion 8, commentary, ¶ 4.

34 *Id*.

35 *See* Michael Schmitt, *France's Major Statement on International Law and Cyber: An Assessment*, Just Security (Sept. 16, 2019), https://www.justsecurity.org/66194/frances-major-statement-on-international-law-and-cyber-an-assessment/ [hereinafter Schmitt, *France's Major Statement*].

robust cyber capabilities simply do not have that luxury. They should be considered specially affected States in the formation of norms relating to the permissibility of collective countermeasures.

Whether State practice regarding collective countermeasures is extensive is difficult to discern, and so far, *opinio juris* is far from uniform, as previously noted. Many of the States that have issued public views on countermeasures, for example, have remained silent on collective countermeasures. It is therefore even more important to remember the role in the evolution of this norm played by States that do not have the independent capacity to engage in countermeasures to respond to malicious cyber operations, as it is precisely these States that would require assistance. States that could be considered specially affected should join Estonia, Costa Rica, and others and make their views known as this norm continues to develop.

3. INTERNATIONAL LAW ON COUNTERMEASURES

What about States like France that argue that collective countermeasures are prohibited? It is generally undisputed that injured States are entitled to take countermeasures. Additionally, shortly after both Estonia and France had declared their respective positions on the legality of collective countermeasures, there was already wide agreement that their legal permissibility was, at a minimum, an open question.[36] To properly assess the legality of collective countermeasures, it is necessary to revisit the history of the debate at the International Law Commission (ILC) and try to pinpoint the reasoning behind such opposition.

A. A Brief History of Countermeasures

The law of State responsibility, under which countermeasures fall, has long been an important focus in the development of international law. It was selected as one of the first fourteen topics for the ILC to address after its creation in 1948.[37] In the following decades, the topic was revisited, and numerous special rapporteurs undertook new readings until, finally, in 2001, under Judge James Crawford, a final version and commentary were issued.[38] As the *Tallinn Manual* notes, the Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA or Articles on State Responsibility) are not binding, but they represent more than fifty years of negotiations between States and have been relied upon in countless international judgments.[39] They should form the starting point of any discussion on countermeasures.

36 Przemysław Roguski, *Collective Countermeasures in Cyberspace—Lex Lata, Progressive Development or a Bad Idea?, in* 2020 12th International Conference on Cyber Conflict 25 (T. Jančárková et al. eds., 2020).

37 James Crawford, *Articles on Responsibility of States for Internationally Wrongful Acts: Introductory Note*, UN Audiovisual Lib. Int'l Law (2012), https://legal.un.org/avl/ha/rsiwa/rsiwa.html.

38 *Id*.

39 Tallinn Manual 2.0, *supra* note 19, at 79, n. 1.

ARSIWA does not include a definition of "countermeasures." However, Article 49 stipulates that they can only be taken in response to an internationally wrongful act to induce compliance by the responsible State; they must be temporary; and they should, as far as possible, allow the injured State to resume performance of the obligation—that is, they should be reversible.[40] Additionally, Article 50 prohibits the use of any countermeasures that would violate human rights law, international humanitarian law (IHL), and peremptory norms, like countermeasures that would constitute a use of force.[41] It is beyond the scope of this work to enter into an exhaustive analysis of the other elements of countermeasures, such as the prior notification or attribution requirements. However, it is important to note that, just as States are shaping the development of international law by advancing their positions on collective countermeasures, they can also shape the requirements of countermeasures as applied in cyberspace to ensure they remain effective.[42]

The question of the permissibility of collective countermeasures was one of the most difficult issues in the final stages of drafting the Articles on State Responsibility.[43] According to Judge Crawford, there was extensive debate, including an initial broad definition of the concept of injured States that included the right of a third State to engage in countermeasures at the injured State's request, much like the system of collective self-defense.[44] States, however, were concerned that this might duplicate work that should take place under Chapter VII of the United Nations Charter such as what measures should be taken to respond to a threat or breach of international peace and security.[45] Other States viewed the law on this question as still developing and felt it was premature to include a definitive assessment.[46] The initially suggested article was thus replaced with a saving clause.[47]

In the end, the ILC decided that, in 2001, the issue of the permissibility of collective countermeasures was in its nascent stage, and thus it chose not to address the question definitively.[48] Article 54 does, however, stipulate that a non-injured State may take "lawful measures" in response to a breach that violates an international obligation

40 Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries, art. 49, *in* Int'l Law Comm'n, Rep. on the Work of Its Fifty-Third Session, 44 U.N. Doc. A/56/10 (2001) [hereinafter ARSIWA].

41 *Id*. art. 50.

42 *See, e.g.*, the positions of the U.S. and the U.K. in *Countermeasures, supra* note 1 (arguing that prior notification is not required when it would render the measures ineffective or expose capabilities of the injured State).

43 James Crawford, *The ILC's Articles on Responsibility of States for Internationally Wrongful Acts: A Retrospect*, 96(4) Am. J. Int'l L. 874, 884–885 (2002) [hereinafter Crawford, *ARSIWA Retrospect*].

44 James Crawford, *State Responsibility*, Max Planck Encyclopedia Int'l L., ¶ 57 (Sept. 2006), https://opil.ouplaw.com/display/10.1093/law:epil/9780199231690/law-9780199231690-e1093 [hereinafter Crawford, *State Responsibility*].

45 *Id*. ¶ 58.

46 *Id*.

47 Crawford, *ARSIWA Retrospect, supra* note 43, at 874, 875.

48 *See* Lahmann, *supra* note 18, at 139; François Delerue, Cyber Operations and International Law 457 (2020).

erga omnes, as defined in Article 48(1).[49] Importantly, even though a majority of States did not think a basis for collective countermeasures could be found in international law at the time, in 2001, the ILC expressly rejected the idea that the only permissible countermeasures were those of a bilateral nature.[50] Finally, the ILC agreed that the issue should be resolved in the future according to changing norms.[51] More than twenty years later, as States continue to publicly express their views, those future developments are happening now.

B. The French Position

France's opposition to collective countermeasures is not new. The country held a similar view during the ARSIWA debates, when it criticized the idea that countermeasures could be taken by non-injured States, even in response to *erga omnes* violations.[52] It felt that third States with a legal interest should be limited to demanding the cessation of the wrongful act in question.[53] That stance appears to have remained unchanged, given France's positions in 2019 and 2023 on the applicability of international law to operations in cyberspace. However, as is common for State positions, no citation is provided to support the assertion that collective countermeasures are "not authorized" under international law. Nor have I been able to find any official or unofficial explanation for the French position.[54]

Nevertheless, a 2017 study on the second *Tallinn Manual* commissioned by the French Ministry of the Armies, the same body that authored France's official position, seems to base the assertion that only victim States may engage in countermeasures on the ICJ's 1986 *Nicaragua* judgment.[55] The study notes that the *Tallinn Manual* experts are divided on the question of collective countermeasures but that a majority of the experts agree with the prohibition as framed by the Court, while only a minority think that a third State may take countermeasures at the request of a victim State.[56] The *Tallinn Manual* indeed reflects this split, observing that a majority of the experts felt that "as set forth in the *Nicaragua* judgment, purported countermeasures taken on behalf of another State are unlawful."[57]

49 ARSIWA, *supra* note 40, art. 54. *See also* James Crawford, State Responsibility: The General Part 66–67 (2013) [hereinafter Crawford, State Responsibility: The General Part].

50 Crawford*, State Responsibility, supra* note 44, at ¶ 58.

51 *Id*.

52 Crawford, State Responsibility: The General Part, *supra* note 49, at 66, 87.

53 *Id*.

54 *Cf.* Przemysław Roguski, *France's Declaration on International Law in Cyberspace: The Law of Peacetime Cyber Operations, Part II*, Opinio Juris (Sept. 24, 2019), http://opiniojuris.org/2019/09/24/frances-declaration-on-international-law-in-cyberspace-the-law-of-peacetime-cyber-operations-part-ii/; Schmitt, *France's Major Statement, supra* note 35; Schmitt & Watts, *supra* note 15.

55 François Delerue, *Analyse du Manuel de Tallinn 2.0 sur le droit international applicable aux cyber-opérations*, CEIS 35 (Nov. 2017), http://francoisdelerue.eu/wp-content/uploads/2020/01/20171129_NP_F-Delerue_Analyse-Manuel-Tallinn-2-0.pdf. See also Delerue, *supra* note 48, at 454–455.

56 Delerue, *supra note 55*, at 35.

57 Tallinn Manual 2.0, *supra* note 19, at 132.

This is an exceedingly narrow reading of the ICJ's judgment in a case where the dispute centers on measures involving the use of force. Nicaragua's complaint, as the name of the case *Case Concerning the Military and Paramilitary Activities in and Against Nicaragua* indicates, mainly relates to "the actual use of force against it" by the United States.[58] The measures engaged in by the United States include training, financing, and equipping an armed group, the *contras*, as well as placing mines and attacking ports, oil installations, and a naval base on Nicaraguan territory.[59] The justification the United States presented for its forcible actions was that it had engaged in collective self-defense to assist El Salvador, Honduras, and Costa Rica after Nicaragua attacked them.[60] The Court rejected the collective self-defense argument upon finding that Nicaragua's actions, while constituting an unlawful use of force, did not amount to an armed attack.[61]

Having rejected the collective self-defense argument, and given the United States's non-participation in the merits phase, the Court felt bound to consider whether the United States's actions could be justified as countermeasures.[62] In the often-cited paragraph 249 of the judgment, the Court recalled that it had already found that use of force short of an armed attack did not justify "collective counter-measures involving the use of force." It added that, even if Nicaragua had violated the principle of non-intervention, only "proportionate counter-measures on the part of the State which had been the victim of these acts" would be justified.[63] It concluded that such a violation imputable to Nicaragua "could not justify counter-measures taken by a third State [the United States], and particularly could not justify intervention involving the use of force."[64] It is usually this language that forms the basis of the view that non-injured States are not permitted to engage in collective countermeasures.

In my view, a more accurate reading of the judgment is that it prohibits countermeasures taken by a third State that "involv[e] the use of force."[65] Here the Court's own characterization is useful. Referring to its analysis, including paragraph 249, the Court observed that it had "disposed of the suggestion of a right to collective countermeasures in [the] face of an armed intervention."[66] In addition, regarding the formation of custom, the views of the parties cannot be disregarded. The Court itself acknowledges that it does not have the "authority to ascribe to States legal views which they do not

58 Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.), Judgment, 1986 I.C.J. Rep. 14, ¶ 227 (June 27).
59 *Id*. ¶¶ 76–122, 227–238.
60 *Id*. ¶¶ 126–130, 161–165, 229–232.
61 *Id*. ¶¶ 232–238.
62 *Id*. ¶¶ 248, 257. *See also* Martin Dawidowicz, Third-Party Countermeasures in International Law 65–66 (2017).
63 *Nicar. v. U.S.*, 1986 I.C.J. ¶ 249.
64 *Id*.
65 *Id*. ¶¶ 211, 249, 252. For an overview of the debate and the opposite conclusion, *see* Dawidowicz, *supra* note 62, at 66–67, n. 163.
66 *Nicar. v. U.S.*, 1986 I.C.J. ¶ 257. *See also id*. ¶¶ 262, 268.

themselves advance."[67] It was not the United States that used the concept of collective countermeasures as a defense but the Court acting under Article 53 of the statute, which requires it to consider all relevant rules in the settlement of the dispute in the event of non-participation by one of the parties.[68] In fact, as the Court recognized, the United States "expressly and solely" justified its actions by reference to the right to engage in collective self-defense in response to an armed attack.[69]

Even assuming the ICJ did find that third-State countermeasures were not authorized, it is essential to recall that the Court was interpreting customary international law as it existed in 1986. ARSIWA—drafted fifteen years later, in 2001—cites the same paragraph that the 2017 French study and the *Tallinn Manual* experts rely on from the *Nicaragua* judgment, paragraph 249, to come to the same conclusion I suggest above: it is forcible countermeasures that are prohibited.[70] The various references to the *Nicaragua* judgment in the final Articles on State Responsibility demonstrate that the ILC thoroughly examined the judgment and yet still concluded that, while at the time there was not sufficient State practice to definitively identify a rule on collective countermeasures, the law as it stood in 2001 did not support the restrictive interpretation that only the injured State could engage in countermeasures.[71] In this sense, it is worth noting that the ICJ itself, in the *Nicaragua* judgment, after recalling the requirements of State practice and *opinio juris* for the formation of custom, observed that "reliance by a State on a novel right or an unprecedented exception to the principle might, if shared in principle by other States, tend towards a modification of customary international law."[72] In the end, even a position based on a more restrictive interpretation of the *Nicaragua* judgment regarding collective countermeasures must adapt to evolving norms.

4. PRACTICAL IMPLICATIONS

The position advanced by States like Estonia and Costa Rica is legally sound, and these States are setting the basis of an emerging legal norm applicable in cyberspace. Their position is also a practical solution for the reality of cyberattacks. As noted previously, most cyberattacks do not reach the armed attack threshold. But this is not a static threshold. States that do not individually have robust cyber capabilities are not likely to accept a position that renders them helpless if they could instead respond with kinetic force already at their disposal. Allowing these States to turn to allies for help short of the use of force could drastically reduce this tendency. Lastly, the very nature of cyberattacks calls for a collective response. Cyberattacks can easily

67 *Id.* ¶ 207.
68 *See id.* ¶¶ 26–31; 266.
69 *Id.* ¶¶ 208, 266.
70 ARSIWA, *supra* note 40, art. 50, commentary, ¶ 5.
71 Crawford, *State Responsibility, supra* note 44, ¶¶ 57–58.
72 *Nicar. v. U.S.*, 1986 I.C.J. ¶ 207.

cause widespread harm beyond the borders of the original victim State. Collective countermeasures could prevent damage from escalating and spreading by allowing States to collaborate rather than wait to be harmed in order to respond individually.

A. Peace and Security

When discussing IHL, it is essential to remember the consequences of armed conflict, such as civilian and combatant deaths, destruction of civilian and military objects, damage to the environment, and the difficulties of post-conflict transition.[73] Scholars have warned that the current restrictions on countermeasures limit a tool that could promote international peace and security, especially compared to more permissive aspects of IHL, and create perverse incentives that push States to expand the law of self-defense to make the use of force a permissible response to more cyberattacks.[74] This tendency is especially likely for States that do not have the independent capacity to respond to subthreshold cyberattacks and that may be tempted to legally justify a resort to the use of force.[75]

Here again, France's policy on the applicability of international law in cyberspace is instructive. France maintains a strict division between cyberattacks that constitute a use of force under Article 2(4) of the UN Charter and those that amount to an armed attack under Article 51.[76] It follows the scale and effects test set out by the ICJ in *Nicaragua* to distinguish unlawful uses of force from armed attacks.[77] At the same time, France's approach to determining when a cyberattack constitutes an unlawful use of force under Article 2(4) and thus legally justifies individual countermeasures goes beyond that adopted by the *Tallinn Manual*.[78] This approach makes sense for a technologically advanced State like France that may wish to respond to a wider variety of cyberattacks with countermeasures while maintaining a high threshold for armed attacks that trigger the law of self-defense.[79] Unlike France, States that cannot independently engage in countermeasures effective against cyberattacks may be tempted to adopt an approach to classifying cyberattacks that triggers the law of self-defense and collective action, as well as eschew what many deem the more restrictive conditions of countermeasures.[80]

73 *See* Naz K. Modirzadeh, *Cut These Words: Passion and International Law of War Scholarship*, 61(1) Harv. Int'l L.J. 1 (2020).

74 *See, e.g.*, Gary Corn & Eric Jensen, *The Use of Force and Cyber Countermeasures*, 32 Temp. Int'l & Comp L.J. 127, 129–132 (2018); Lahmann, *supra* note 18, at 141–146.

75 Corn & Jensen, *supra* note 74, no 130.

76 *France 2023 Position on International Law Applicable to Cyber Operations, supra* note 5, at 6, 8–10.

77 *Id.* at 8.

78 Schmitt, *France's Major Statement, supra* note 35; Roguski, *supra* note 54.

79 *See* Schmitt, *France's Major Statement, supra* note 35.

80 *See, e.g.*, Russel Buchan, *Non-Forcible Measures and the Law of Self-Defence*, 72(1) Int'l & Comp. L.Q. 1, 2–3 (2023) (arguing that self-defense is a "general right" in international law, not an exception to the prohibition on the threat or use of force, and can therefore justify "all measures necessary" against an armed attack, forcible or non-forcible).

B. Collective Action for Widespread Harm

Cyberattacks also have the capacity to spread quickly and cause widespread damage beyond any physical borders. Scholars have highlighted that the very nature of malicious cyber operations and their potential for extensive effects makes collective action an especially well-suited solution.[81] Take, for example, the massive damage wrought by the NotPetya attack in 2017. This started as an attack on Ukrainian networks that was made to look like a ransomware attack so that it would be treated not as a geopolitical attack but as cybercrime. The malware then quickly spread to the United States, France, the United Kingdom, and other countries, causing an estimated US$10 billion in damages.[82] The United States and the United Kingdom have both attributed the attack to Russia, which at the time was involved in an armed conflict with Ukraine that began in 2014.[83]

While it is impossible to be certain, it is not hard to imagine that if States like the United States and the United Kingdom had been allowed to assist Ukraine by engaging in collective countermeasures, once they attributed the attack to Russia, they could have reduced the effects and reach of the damage of the NotPetya attack. Possible countermeasures non-injured States could carry out at the request of victim States include active cyber defense practices like hack-backs, where a State takes proactive action against the source of the malicious cyber operation.[84] Given the realities of hybrid warfare today, collective countermeasures could protect even States that do possess robust independent cyber capabilities but may require assistance from third States if involved in an armed conflict that stretches their available resources.

5. CONCLUSION

We, as members of the international community, should ask ourselves whether we feel comfortable telling a State that would be legally permitted to engage in countermeasures individually that, because it does not possess the resources to do so on its own, or to pay private actors instead, it simply cannot. The reality is that when it comes to cyber capabilities, as with so many other resources, there is a deep imbalance between States such as, for instance, Russia and Estonia. It is critical for small and developing States, especially those in the global majority, to have access to effective methods of responding to malicious cyber operations. Working together would allow these States to contend with the power imbalance that exists in the current

81 Corn & Jensen, *supra* note 74, at 130.

82 *See* Andy Greenberg, *The Untold Story of NotPetya, the Most Devastating Cyberattack in History*, Wired (Aug. 21, 2018), https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/; Lahmann, *supra* note 18, at 12.

83 *UK and US Blame Russia for "malicious" NotPetya Cyber-attack*, BBC News (Feb. 15, 2018), https://www.bbc.com/news/uk-politics-43062113.

84 Lahmann, *supra* note 18, at 124–128; Tallinn Manual 2.0, *supra* note 19, at 563 (on active cyber defense), 565 (on hack backs); Roguski, *supra* note 36, at 40–41.

international order.[85] The law as it currently stands is unsettled on the permissibility of collective countermeasures, a tool that not only could help States that require assistance to effectively respond to cyberattacks but also could prevent recourse to the use of force by States that see no other option. Estonia opened the door five years ago, and since then, several other States that have been victims of cyberattacks, like Costa Rica and Ireland, have walked through. As we face the nascent stage of a developing norm of international law applicable to cyberspace, it is time for States, particularly specially affected States that may require assistance, to make their views known.

ACKNOWLEDGMENTS

I would like to thank the reviewers for their engaging comments; they made this a better paper. Thanks also to my friends and family, who always provide a sounding board, as well as to my colleagues in the Atlantic Council for their support, especially Franklin D. Kramer for his thoughtful feedback and suggestions. All errors are my own.

85 *See, e.g.*, Oonah Hathaway, Maggie Mills & Thomas Poston, *"The Emergence of Collective Countermeasures,"* Arts. War (Nov. 1, 2023), https://lieber.westpoint.edu/emergence-collective-countermeasures/.

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

Anti-Satellite Weapons and Self-Defence: Law and Limitations

Chris O'Meara
Senior Lecturer in Law
Law School
University of Exeter
Exeter, United Kingdom
c.omeara@exeter.ac.uk

Abstract: Space is an increasingly militarized domain with the potential to be a source and place of armed conflict. In recent years, tests of anti-satellite (ASAT) weapons capable of neutralizing civilian and military satellites have fuelled fears of warfare in that domain. Satellites are potentially attractive targets during armed conflict, making ASAT weapons central to assessing the threat environment in space. Space debris resulting from ASAT weapon use is of particular concern, as it threatens other satellites in orbit, many of which underpin the operation of human societies and global economies. Although states recognize this threat, attempts at weapons control have failed. Instead, we must look to existing international law that governs military activities in space, including in the cyber domain. Yet, how the *jus ad bellum* (JAB), which regulates state uses of force, applies to ASAT weapons has received little attention. This is despite state assertions of their right to act in self-defence in space.

This paper argues that JAB regulation of ASAT technologies addresses state concerns regarding protecting space assets and avoiding conflict in space. This author contends that states acting defensively in space are restricted in their choice of targets by the requirements of JAB necessity and proportionality, which protect civilians and the interests of other states. While defensive acts, including cyber operations, that do not cause space debris are most likely to be JAB-compliant, this is not guaranteed. Military actions of any kind against mixed-use or multi-user satellites raise particular concerns for JAB proportionality due to the potential resulting harm to civilians and to the interests of other states. A clearer understanding of how the JAB regulates ASAT weapons helps decision makers avoid lawful acts of self-defence being characterized

as unlawful uses of force. Adherence to these JAB rules ultimately helps secure international peace and security on Earth and beyond Earth's atmosphere.

Keywords: *ASAT weapons, jus ad bellum, necessity and proportionality, self-defence, space*

1. INTRODUCTION

Space activities underpin all instruments of national power,[1] with states increasingly considering space as an integrated part of their national security. NATO, for example, recognizes space as an operational domain, alongside air, land, sea, and cyberspace.[2] Space is also a contested domain, with the potential to be a source and place of armed conflict. Unease over the 'weaponization' of space[3] is accordingly at the top of the international agenda, with the UN General Assembly (UNGA) consistently emphasizing the need for international cooperation on the peaceful uses of outer space and expressing serious concern about an arms race in that domain.[4] There exists particular unease over states developing counterspace weapons that threaten access to and freedom to operate in space.[5] Although no state has yet used such a weapon against another state's satellite,[6] given their importance to military operations, satellites might be considered attractive targets in armed conflict. This fact is evidenced by the testing of offensive and defensive anti-satellite (ASAT) weapons capable of disrupting or destroying both civilian/commercial and military satellites.

The call for legal regulation of ASAT weapons is urgent, given the physics of space and the potential enduring effects of space debris that might result from ASAT weapon use.[7] Space debris imperils other satellites in orbit, many of which are fundamental to the operation of human societies and global economies. States view space debris as a significant threat to the space environment, with the intentional destruction of satellites exacerbating the threat.[8] Yet, multilateral attempts to restrain the

1 Kari A Bingen, Kaitlyn Johnson, Makena Young, and John Raymond, *Space Threat Assessment 2023* (Center for Strategic and International Studies, April 2023) 1 (CSIS, *Space Threat Assessment 2023*).

2 NATO, 'NATO's Overarching Space Policy' (17 January 2022) <www.nato.int/cps/en/natohq/official_texts_190862.htm> accessed 1 March 2024.

3 See Jinyuan Su, 'Use of Outer Space for Peaceful Purposes: Non-Militarization, Non-Aggression and Prevention of Weaponization' 36(1) Journal of Space Law (2010) 253.

4 For example, UNGA Res 55/122 (27 February 2001) UN Doc A/RES/55/122; UNGA Res 72/78 (14 December 2017) UN Doc A/RES/72/78; UNGA Res 77/41 (12 December 2022) UN Doc A/RES/77/41.

5 NATO's Overarching Space Policy (n 2) para 2.

6 CSIS, *Space Threat Assessment 2023* (n 1) 4.

7 Orbital debris is any human-made object in orbit around the Earth that no longer serves any useful purpose. Depending on the orbit, space debris may endure for hundreds of years or more. NASA Orbital Debris Program Office, 'Frequently Asked Questions' <https://orbitaldebris.jsc.nasa.gov/faq/> accessed 1 March 2024.

8 UNGA Res 77/41 (n 4).

escalating weaponization of space have failed.[9] Legally regulating ASAT weapons consequentially relies on bodies of international law that were not originally designed for space but naturally pertain to ASAT weapon use. To date, much of the focus regarding conflict in space has understandably been on international humanitarian law (IHL).[10] Meanwhile, the regulatory potential of the *jus ad bellum* (JAB), which governs when states may use force in their international relations, has received little attention.[11] This is an opportunity lost. This paper addresses this gap by investigating how the JAB restricts the use of ASAT weapons and responds directly to state concerns over their possible use, including in self-defence.

Section 2 begins by explaining the different types of ASAT weapons and the associated international law, including the application of the JAB. Sections 3 and 4 explore how states exercising their right of self-defence in space are restricted by the JAB. Section 3 emphasizes that, despite the potential military advantage that might be gained by targeting satellites, JAB necessity restricts the target options available to a defending state.[12] Even if IHL and JAB necessity do not prevent a state from targeting a satellite, Section 4 illustrates how the operation of JAB proportionality might nevertheless prohibit ASAT weapon use because of the potential resulting harm to civilians and the interests of other states. Applying these JAB rules has implications at both the strategic level (regarding how states develop space-related policies and ASAT technologies) and the operational and tactical levels (in terms of how military planners execute military operations in space). A clearer understanding of how the JAB rules apply in space helps decision makers avoid acts of self-defence being characterized as unlawful uses of force. Ultimately, JAB regulation of ASAT technologies addresses state concerns regarding protecting their space assets while also helping to avoid conflict and the escalation of conflict in space. Adherence to the JAB rules promotes and helps to secure international peace and security on Earth and beyond the Earth's atmosphere.

2. ASAT WEAPONS AND INTERNATIONAL LAW

The development of counterspace weapons that can disrupt, degrade, or destroy satellites and related infrastructure has a long history, going back to the dawn of the space age.[13] The US, China, and Russia currently possess the most advanced

9 Paul B Larsen, 'Outer Space Arms Control: Can the USA, Russia and China Make This Happen' (2018) 23 J Confl Secur Law 137.

10 For example, Michael N Schmitt, 'International Law and Military Operations in Space' 10 UNYB (2006) 89, 114–24; Jack Mawdsley, 'Applying Core Principles of International Humanitarian Law to Military Operations in Space' (2020) 25 J Confl Secur Law 263; Hitoshi Nasu, 'Targeting a Satellite: Contrasting Considerations between the *Jus Ad Bellum* and the *Jus in Bello'* (2022) 99 Int'l L Stud 142.

11 An exception is Fabio Tronchetti, 'The Right of Self-Defence in Outer Space: An Appraisal' (2014) 63 ZLW 92.

12 A 'defending state' is a state that is, or claims to be, the victim of an armed attack.

13 See Center for Strategic and International Studies, 'Counterspace Timeline, 1959–2022' (31 March 2021) <https://aerospace.csis.org/counterspace-timeline/> accessed 1 March 2024.

ASAT technologies, although other states possess counterspace capabilities.[14] ASAT weapons can be placed into four broad categories.[15] The first category is kinetic physical ASAT weapons, which comprise anti-satellite missiles and other methods of physical kinetic attacks directed against satellites. Such attacks may be launched from the ground (direct ascent ASAT weapons) or from space (co-orbital ASAT weapons). The US, China, Russia, and India have all demonstrated kinetic ASAT capabilities.[16] Non-kinetic physical ASAT weapons, meanwhile, have reversible or permanent physical effects on satellites or other space systems but make no physical contact with them. They include directed energy weapons launched from other satellites, or from land, sea, or airborne weapons platforms on Earth. A third type is electronic ASAT weapons, such as jamming devices that interfere with the transmission of signals to and from satellites and spoofing devices that can falsify signals. The final category comprises cyber attacks that target data and the systems that use data. Such attacks may be used to monitor data, or to intercept, falsify, or corrupt it, and may be temporary or permanent.

Contemporary international space law that governs military activities in space is centred on the Outer Space Treaty of 1967 (OST)[17] and, to a much lesser extent, the Moon Agreement of 1979 (Moon Agreement).[18] Neither treaty prevents ASAT weapons from being used in lawful acts of self-defence.[19] Efforts by the UN to forestall the weaponization of space and to preserve it for peaceful purposes continue but have not yet borne fruit.[20] A comprehensive ban is unlikely in the present geo-political climate.[21] Instead, guiding principles governing space activities and other 'soft law' efforts have been pursued by the UN[22] and the EU,[23] while academic projects like the

14 See Office of the Director of National Intelligence, '2023 Annual Threat Assessment of the U.S. Intelligence Community' (8 March 2023) 8, 15. See also CSIS, *Space Threat Assessment 2023* (n 1).

15 This paper adopts the commonly used categorization of counterspace weapons set out in CSIS, *Space Threat Assessment 2023* (n 1) 3–7.

16 CSIS, *Space Threat Assessment 2023* (n 1) 11, 14, 23; Mawdsley (n 10) 279.

17 1967 Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies, 610 UNTS (OST) 205.

18 1979 Agreement Governing the Activities of States on the Moon and Other Celestial Bodies, 1363 UNTS 3. The Moon Agreement has received a very limited number of signatories and ratifications, with a notable absence of major spacefaring powers. It is not, therefore, a significant source of space law. See UN Office for Outer Space Affairs, 'Status of International Agreements Relating to Activities in Outer Space' <https://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties/status/index.html> accessed 1 March 2024.

19 See n 29.

20 See UN Office for Outer Space Affairs <www.unoosa.org/oosa/en/aboutus/index.html> accessed 1 March 2024; the work of the Committee on the Peaceful Uses of Outer Space, including delegate statements to the Sixty-Sixth Session (31 May–9 June 2023) <www.unoosa.org/oosa/en/ourwork/copuos/2023/statements.html> accessed 1 March 2024. See also the work of the UN Conference on Disarmament <https://disarmament.unoda.org/conference-on-disarmament/> accessed 1 March 2024.

21 See further Shang Kuan, 'Legality of the Deployment of Anti-Satellite Weapons in Earth Orbit: Present and Future' (2010) 36 J Space L 207, 227–30.

22 See UN Office for Outer Space Affairs, 'Space Law Treaties and Principles' <https://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties.html> accessed 1 March 2024.

23 European Union, 'Draft International Code of Conduct for Outer Space Activities' (Version 31 March 2014) <www.eeas.europa.eu/sites/default/files/space_code_conduct_draft_vers_31-march-2014_en.pdf> accessed 1 March 2024.

Woomera and MILAMOS manuals seek to clarify rules pertaining to the use of ASAT weapons.[24]

In the absence of a treaty banning ASAT weapon testing and use, all military activities in space, including the use of ASAT weapons, must nevertheless comply with general public international law, including the rules of the UN Charter.[25] Although the focus to date has naturally been on the detailed targeting rules of IHL, the JAB, grounded in the UN Charter and customary international law, also applies to ASAT weapon use.[26] Under Article 2(4) of the UN Charter, states are prohibited from threatening or using force in their international relations,[27] subject to two recognized exceptions, being force authorized by the UN Security Council (UNSC) under Chapter VII of the UN Charter and force used in self-defence pursuant to Article 51 and customary international law.[28] Accordingly, absent UNSC authorization, any actions against satellites that constitute a threat or use of force require justification as lawful acts of self-defence or they will contravene Article 2(4) as unlawful uses of force.

3. NECESSITY IN OUTER SPACE

A state's right of self-defence arises when an armed attack occurs.[29] JAB necessity and proportionality then apply to the entirety of a defensive military operation to condition the exercise of that right so that force is contained and confined purely to the defensive.[30] They are requirements of customary international law that must be strictly adhered to in order for acts of self-defence to be considered lawful.[31] Necessity first stipulates that defensive force be a measure of last resort, where peaceful alternatives are unavailable or unfeasible and/or, on their own, will be ineffective to halt, repel,

24 Woomera Manual on the International Law of Military Space Activities and Operations <https://law.adelaide.edu.au/woomera/> accessed 1 March 2024; The McGill Manual on International Law Applicable to Military Uses of Outer Space (MILAMOS) <www.mcgill.ca/milamos/> accessed 1 March 2024.

25 OST (n 17) art III. The exceptions are rules that are domain-specific, geographically constrained, or otherwise incompatible with the space environment. Kubo Mačák, 'Military Space Operations', in Sergey Sayapin and others (eds), *International Conflict and Security Law: A Research Handbook* (Springer 2022) 399, 406. See further Frans G von der Dunk, 'Armed Conflicts in Outer Space: Which Law Applies?' (2021) 97 Int'l L Stud 188.

26 The UN Charter applies 'to any use of force, regardless of the weapons employed'. *Legality of the Threat or Use of Nuclear Weapons* (Advisory Opinion) [1996] ICJ Rep 226, para 39.

27 1945 Charter of the United Nations, 892 UNTS 119 (UN Charter), art 2(4).

28 Art 51 recognizes a state's inherent right of individual or collective self-defence if an armed attack occurs. States assert their right to act in self-defence in space and while not universally accepted, the dominant view is that states may lawfully exercise that right in space and the JAB applies there to condition the exercise of that right. Tronchetti (n 11) 104–7; Nasu (n 10) 153; Von der Dunk (n 25) 199, 208–9; Mačák (n 25) 407. See also NATO's Overarching Space Policy (n 2) para 12; MILAMOS (n 24) rule 152.

29 UN Charter (n 27) art 51.

30 Necessity and proportionality apply on an ongoing basis, throughout the duration of an armed conflict prompted by self-defence. Judith Gardam, *Necessity Proportionality and the Use of Force by States* (CUP 2004) 155–56; Tom Ruys, *'Armed Attack' and Article 51 of the UN Charter: Evolutions in Customary Law and Practice* (CUP 2010) 124.

31 *Military and Paramilitary Activities in and Against Nicaragua (Nicaragua v United States)* (Judgment) [1986] ICJ Rep 14, para 176.

or (if some form of anticipatory self-defence is accepted)[32] prevent an armed attack. Force used in self-defence must be the only reasonable choice of means available to the defending state in the circumstances.[33] Any acts of self-defence, in space or on Earth, must surpass this initial hurdle to be considered lawful under the JAB. Any use of force that is unnecessary is unlawful. For defending states considering using ASAT weapons, therefore, the first question is whether options not involving military force are practical and will likely be effective to counter an armed attack, or have a reasonable chance of doing so.[34] In addition to UNSC responses, diplomacy, countermeasures, or military action falling below the threshold of a use of force are obvious alternatives. The latter category might include limited cyber operations that temporarily disable a satellite but do not cause it to collide with other space objects or otherwise create debris. Only when such non-forceful alternatives, on their own, are insufficient to respond to an armed attack does the necessity of using force in self-defence arise. Unless and until this requirement is satisfied, ASAT weapons that comprise a use of force may not be used.

Where reasonable alternatives are unavailable or will be ineffective, targeting satellites in self-defence using ASAT weapons might be considered an attractive option for defending states.[35] Satellites are an integral part of modern warfare, providing precise navigation, furnishing real-time targeting and weather data, allowing instantaneous global communications, warning of possible missile threats, collecting intelligence, and carrying out surveillance and reconnaissance.[36] Satellites might also be the source of an armed attack in space. As such, the defensive advantage of neutralizing a satellite that supports an adversary's aggressive behaviour seems readily apparent. Yet, even if the *prima facie* necessity is established for a state to resort to force in self-defence in some form, it does not follow that satellites automatically become fair game. To be lawful, targeting satellites must comply with the further requirement of JAB necessity that confines defensive responses to targets that serve a defensive purpose.[37] For policymakers and military planners, therefore, JAB necessity constitutes a further hurdle to overcome, in addition to complying with IHL targeting rules that also govern whether or not satellites are targetable. Under IHL, satellites may be targeted if they constitute 'military objectives' and IHL proportionality limitations are adhered to. 'Civilian objects' may not be directly targeted.[38] In addition, JAB necessity restricts

32 See Chris O'Meara, 'Reconceptualising the Right of Self-Defence Against 'Imminent' Armed Attacks' (2022) 9(2) J Use Force Int Law 278.

33 Chris O'Meara, *Necessity and Proportionality and the Right of Self-Defence in International Law* (OUP 2021) 38–42.

34 Elizabeth Wilmshurst, 'The Chatham House Principles of International Law on the Use of Force in Self-Defence' (2006) 55(4) ICLQ 963, 967.

35 'Targeting' a satellite involves engagement or action to alter or neutralize the function it performs for the adversary. United States Chairman of the Joint Chiefs of Staff, Joint Publication 3-60, *Joint Targeting* (31 January 2013) I-1. This may mean physically damaging or destroying a satellite or otherwise permanently or temporarily disabling or neutralizing it.

36 Schmitt (n 10) 90.

37 See O'Meara (n 33) 84–93.

38 1977 Protocol Additional to the Geneva Conventions, 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts, 1125 UNTS 3 (AP I) arts 48, 51(2), 52(2), 57(2).

these targeting decisions to ensure they remain defensive.[39] JAB necessity imposes an additional and separate targeting obligation that operates cumulatively and in parallel to these IHL rules, requiring that defensive force be (i) directed against the source of the armed attack[40] and (ii) limited to military targets connected with that armed attack.[41]

That JAB necessity requires self-defence to be connected to an armed attack means that it is only necessary to target military assets that belong to the authors of that attack. Yet, the challenges associated with legally attributing armed attacks to their author[42] are potentially exacerbated in space. An incorrect assessment of attribution might result in military action being characterized as an unlawful armed attack because it is directed at the wrong object. As well as breaching Article 2(4) of the UN Charter, if rising to the level of an armed attack, misdirected action risks a military response by the target state in self-defence. Tracing the use of ASAT weapons to their author depends heavily on the technology employed. Direct ascent ASAT missile attacks against satellites are more easily attributed than other counterspace weapons because launches from Earth are detectable and their effects can create identifiable orbital debris.[43] Beyond that method of attack, however, identifying aggressive acts in space is generally challenging. Other technologies, such as non-kinetic directed energy ASAT weapons, electronic ASAT weapons, and cyber operations, are often much less visible and are consequently more difficult to attribute to an aggressor.[44] Moreover, not all threats to satellites come from states. Acts like signal jamming and cyber attacks might also be carried out by non-state actors, including terrorist organizations.[45] This possibility further complicates the threat assessment and related response. Ultimately, on a good faith assessment, if there is no reasonable and objective basis for concluding that a particular state is the author of an armed attack, JAB necessity precludes the targeting of that state's satellites in response, together with any other asset belonging to that state. For attacks not involving direct ascent ASAT missiles, this is likely to be a high evidential burden to meet.

The nature and composition of satellites also factor heavily in whether satellites may be lawfully targeted in self-defence. Satellites can be purely civilian or commercial in nature, meaning they are not lawful targets for the purposes of JAB necessity. Alternatively, it may be established that a satellite registered to an aggressor state clearly serves a military purpose, is wholly owned or used by it, and is factually connected with the armed attack being defended against. If so, JAB necessity imposes

39 *Case Concerning Oil Platforms (Iran v United States)* (Judgment) [2003] ICJ Rep 161, paras 51, 74–77; O'Meara (n 33) 84–93.
40 Ruys (n 30) 108–9.
41 See further below in this section regarding the military nature of targets.
42 See Ruys (n 30) 368–510.
43 United States Defense Intelligence Agency, *2022 Challenges to Security in Space*, 46 <www.dia.mil/Portals/110/Documents/News/Military_Power_Publications/Challenges_Security_Space_2022.pdf> accessed 1 March 2024.
44 CSIS, *Space Threat Assessment 2023* (n 1) 4–5.
45 NATO's Overarching Space Policy (n 2) para 4.

no obvious restrictions on targeting it.[46] Yet, verifying the military nature of a satellite is generally difficult. Moreover, other satellites may not have one sole function. They may be dual-use, hosting multiple payloads, some serving civilian or commercial purposes, and others having military functions. Payloads themselves, notably providing communication and navigation services, may also be dual-use in that they serve both civilian/commercial and military clients.[47] SpaceX's Starlink service, which provides internet access to support Ukraine's self-defence against Russia's ongoing invasion, is a good example[48] since civilians have likewise relied heavily on the service.[49] In addition to mixed civilian and military use, Starlink demonstrates how satellites may not be exclusively owned, operated, or used by one state but rather by a multiplicity of states, international organizations, and/or private entities.[50]

Mixed-use and multi-owner/user satellites pose significant challenges for military planners. Even if connected with an armed attack, the fact that a satellite may not be exclusively military in nature and/or may not be owned, operated, or used solely by the aggressor state logically limits actions against it. At a minimum, JAB necessity restricts which type of ASAT weapon may be used to target a satellite. Direct ascent or co-orbital kinetic physical ASAT weapons that are not capable of being directed at particular payloads and cannot avoid harm to other payloads on the same satellite are ruled out by this requirement. Their effects cannot be limited solely to military objects connected to the armed attack. Anti-satellite missiles are the obvious example, as they are likely to destroy or damage the entire satellite and all payloads without distinction. Other types of ASAT weapons, however, do have the potential to comply with the rule that only the aggressor and not innocent third parties are targeted in self-defence, depending on how they are used. Such weapons might include co-orbital robotic arm technologies, non-kinetic ASAT weapons like electronic devices that jam specific military signals, or cyber operations that can be specifically targeted and their effects contained. However, limiting the effects of these types of weapons is not guaranteed due to the mixed-use or multi-user nature of such satellites, as well as the inherent difficulties of operating in space and any action the victim of an intended attack might take to thwart it, which might result in kinetic impacts.

46 Targeting may breach the requirements of JAB proportionality, however. See Section 4.

47 Nasu (n 10) 143–44.

48 'How Elon Musk's Satellites Have Saved Ukraine and Changed Warfare', *The Economist* (5 January 2023) <www.economist.com/briefing/2023/01/05/how-elon-musks-satellites-have-saved-ukraine-and-changed-warfare?utm_medium=cpc.adword.pd&utm_source=google&ppccampaignID=18156330227&ppcadID=&utm_campaign=a.22brand_pmax&utm_content=conversion.direct-response.anonymous&gclid=Cj0KCQjw4s-kBhDqARIsAN-ipH0xw65Ux5B1mojK1AQkrW3EOl33GC_qyAHXbu3XGp9MuSEKM7VE1xoaAiMuEALw_wcB&gclsrc=aw.ds> accessed 1 March 2024.

49 Daryna Antoniuk, 'How Elon Musk's Starlink Satellite Internet Keeps Ukraine Online', *The Kyiv Independent* (3 September 2022) <https://kyivindependent.com/how-elon-musks-starlink-satellite-internet-keeps-ukraine-online/> accessed 21 March 2024.

50 Intelsat and Immarsat are other obvious examples. See PJ Blount, 'Targeting in Outer Space: Legal Aspects of Operational Military Actions in Space' *Harvard National Security Journal, Features, Online Edition* (25 November 2012) <https://harvardnsj.org/2012/11/25/targeting-in-outer-space-legal-aspects-of-operational-military-actions-in-space/> accessed 1 March 2024.

Given these risks, defending states should avoid targeting mixed-use and multi-owner/user satellites entirely. Only targeting satellites that are (i) solely owned, operated, or used by the aggressor state, (ii) exclusively military, and (iii) factually connected with the armed attack, will likely comply with JAB necessity. However, given the potential effects of damaging or destroying any type of satellite, even if the JAB necessity (and IHL) requirements are capable of being satisfied, JAB proportionality is likely to act as a strong limitation on whether or not a satellite may lawfully be targeted in self-defence and might even prohibit the targeting of certain satellites entirely.

4. PROPORTIONALITY IN OUTER SPACE

If self-defence is necessary, JAB proportionality proceeds to restrict how much total force states may use in military operations to achieve a legitimate defensive purpose. It permits states to effectively defend themselves but requires that states do no more than that.[51] Unlike IHL proportionality, which operates at an operational level of decision-making to minimize collateral harm to civilians resulting from individual planned attacks,[52] JAB proportionality operates at the strategic level to limit a defending state's total military response viewed as a whole. It is, therefore, a prohibition against excessive overall military reactions by states that undertake necessary acts of self-defence. More specifically, JAB proportionality requires states to balance their defensive force and its outcomes primarily against the legitimate defensive purpose. It also requires that defensive operations do not have undue negative impacts on civilians and on the interests of other states and the international community more broadly.[53]

JAB proportionality has particular significance in space due to its physical characteristics. The dangers associated with the possible use of prohibited nuclear ASAT weapons in this domain are manifest,[54] and the risk of enduring space debris caused by damage to or destruction of satellites by conventional weapons is also a factor peculiar to the space environment. Moreover, the risk is growing. Space is an increasingly congested domain. More than 6,700 satellites currently orbit the

51 O'Meara (n 33) 97–100.

52 IHL proportionality requires an assessment of whether expected civilian loss or injury and damage to civilian objects that result from a particular planned attack would be excessive in relation to the concrete and direct military advantage anticipated. AP I (n 38) art 51(5)(b).

53 See O'Meara (n 33) 100–25, 146–55.

54 The placing of nuclear weapons and other weapons of mass destruction in orbit, on celestial bodies, and otherwise stationing such weapons in outer space is prohibited. OST (n 17) art IV; 1963 Treaty Banning Nuclear Weapon Tests in the Atmosphere, in Outer Space and Under Water, 480 UNTS 43. On the possible devastating effects of detonating nuclear weapons in space, see David Wright, Laura Grego, Lisbeth Gronlund, *The Physics of Space Security: A Reference Manual* (2005) 138–39; Charlie JP Bennett, 'Nuclear Space-Based ASAT Weapons – A Brief International Legal Perspective' (27 February 2024) *EJIL: Talk!* <https://www.ejiltalk.org/nuclear-space-based-asat-weapons-a-brief-international-legal-perspective/?utm_source=mailpoet&utm_medium=email&utm_campaign=ejil-talk-newsletter-post-title_2> accessed 1 March 2024.

Earth,[55] with one estimate predicting 24,500 satellites in orbit by the end of 2031.[56] Debris-creating defensive ASAT weapon use could accordingly have long-lasting and unforeseen consequences for the rights and interests of many spacefaring actors. Space debris does not discriminate,[57] so the risk of collision with debris 'is to all civilian, commercial, and government satellites of all nations'.[58] States, therefore, view space debris as 'the most significant threat to the space environment', with 'the intentional destruction of satellites using kinetic force as exacerbating such threats'.[59] Any or all of their satellites, together with any vital services that rely on them, could be affected to varying degrees by ASAT weapons use. Adherence to JAB proportionality addresses this concern.

A. Civilian Harm

On Earth, 'it is the strategic impact of large-scale civilian casualties and damage that appears to influence what might constitute a disproportionate exercise of the right to self-defence by a State'.[60] That civilian harm stands as the clearest indicator of JAB disproportionality is clearly reflected in the practice of states.[61] Of greatest significance for present purposes is the potential harm to satellites owned and operated by civilians (including corporations), together with the effects on Earth of damaging or destroying satellites that serve civilian populations. Such consequential civilian harm may result from a satellite being targeted directly by ASAT weapons or because satellites are damaged or destroyed by space debris that has resulted from targeting other satellites. Given the potential enduring nature of debris clouds, any number of satellites belonging to civilians or serving civilian needs are put at risk. That risk is hugely significant, given how central satellites are to human societies and global economies. Even temporary disruption to satellites that serve these vital civilian needs may have effects stretching and enduring well beyond the use of the ASAT weapon.

An obvious example of the significance of this risk is any harm caused to the American Global Positioning System (GPS)[62] and its equivalents,[63] which provide military and civilian users with global positioning, navigation, and timing (PNT) services. PNT services are indispensable to the functioning of modern civilizations. Agriculture, transport networks (including global aviation), financial markets, banking systems,

55 Union of Concerned Scientists, 'Union of Concerned Scientists Satellite Database' (1 May 2022) <www.ucsusa.org/resources/satellite-database> accessed 1 March 2024.

56 Euroconsult, 'Satellite Demand to Quadruple over the next Decade' (12 December 2022) <www.euroconsult-ec.com/press-release/satellite-demand-to-quadruple-over-the-next-decade/> accessed 1 March 2024.

57 Von der Dunk (n 25) 227.

58 *2022 Challenges to Security in Space* (n 43) 37.

59 UNGA Res 77/41 (n 4).

60 Kenneth Watkin, *Fighting at the Legal Boundaries: Controlling the Use of Force in Contemporary Conflict* (OUP 2016) 62.

61 See O'Meara (n 33) 139–46.

62 See 'GPS: The Global Positioning System' (*GPS.gov*) <www.gps.gov/> accessed 1 March 2024.

63 For example, the Russian Global Navigation Satellite System (GLONASS), China's BeiDou Navigation Satellite System (BDS), the EU's Galileo global navigation satellite system, India's NavIC system, and Japan's Quasi-Zenith Satellite System (QZSS).

logistics, communications systems, critical infrastructure (such as power grids), emergency services, environmental protection, disaster surveillance, military operations, and the preservation of national security more generally, all rely on PNT technology.[64] Even a temporary and reversible disruption to PNT services could have disastrous consequences for millions of civilians who rely on them on Earth. The effects might be economic, caused by havoc wrought on financial markets. In this regard, the head of UK Space Command has noted how Russia could potentially use jamming satellites that could 'cut off the UK from the outside world'.[65] The effects could equally be physical, for example, because emergency services or disaster relief teams are unable to respond, aircraft cannot fly safely or other transport systems cannot function properly, agricultural production is disrupted, and so forth. The International Committee of the Red Cross (ICRC), among others, has voiced its concern about such potential severe human costs,[66] with civilian injury or death being readily foreseeable in many instances.

Given the importance of PNT services to life on Earth, the potential effects on individuals and on human society caused by ASAT weapons are, in many respects, unforeseeable and unquantifiable. Accordingly, given the requirement to minimize collateral civilian harm, JAB proportionality arguably rules out the direct targeting of satellites that provide PNT and equivalent essential services, as well as the targeting of other satellites (such as those in proximate orbits) that put such essential services at risk because of resulting debris. This JAB prohibition operates separately and in addition to consideration of civilian harm for the purpose of compliance with IHL proportionality.[67] Even the use of non-kinetic ASAT weapons, such as cyber attacks, to disrupt essential services temporarily seems impossible to justify under the JAB. This is so despite a functional link between a satellite and an armed attack that might satisfy JAB necessity. For JAB proportionality, the potential repercussions of these acts on Earth are too varied and potentially too significant to evaluate in any meaningful way that might justify the pursuit of a defensive purpose. The possible repercussions also mean that deploying ASAT weapons, including cyber attacks, against PNT and other essential services that *prima facie* fall below the threshold of a use of force is also a risky strategy. This is because the resulting scale and effects might mean that the threshold of violence is eventually crossed and the JAB requirements that states wished to avoid nevertheless end up applying.

64 'GPS Applications' (*GPS.gov*) <www.gps.gov/applications/> accessed 1 March 2024.
65 George Grylls, 'China "Will Drill Moon for Minerals"' *The Times* (1 July 2023).
66 ICRC, 'The Potential Human Cost of the Use of Weapons in Outer Space and the Protection Afforded by International Humanitarian Law' (April 2021) <https://www.icrc.org/en/document/potential-human-cost-outer-space-weaponization-ihl-protection> accessed 1 March 2024.
67 See O'Meara (n 33) 155–61. IHL proportionality assessments might, or might not, also rule out such acts of targeting, but this is a distinct legal question.

The only possible exception to these conclusions, as with the use of nuclear weapons, is greater freedom to use ASAT weapons where the survival of the state is at stake.[68] Beyond such extreme and unusual circumstances, however, JAB proportionality acts as a significant limitation on defensive action in space. Where satellites that provide non-essential services are targeted and/or where the scale and effects of targeting are in fact limited, the risk is consequentially less. Context is determinative, however, and the wider effects of such targeting must be considered. States must account for impacts on civilians resulting from harm to any kind of satellite, even those not providing essential services. JAB proportionality requires decision makers to consider very carefully the possible resulting consequences in space and on Earth, in each case so as to minimize collateral civilian harm.

B. Third-Party Rights and Interests

Beyond civilian harm, state interests also factor in the JAB proportionality assessment. JAB proportionality requires that the legally protected rights of other states must not be unduly harmed when defending states use ASAT weapons.[69] In addition to potentially breaching the obligation of due regard under Article IX of the OST,[70] excessive harm to such third-party rights risks defensive action being deemed disproportionate under the JAB. Although this assessment is largely fact-dependent, the risk of enduring space debris caused by ASAT weapon use clearly poses a direct threat to satellites owned or operated by third states and to satellite-provided services on which such states rely (including essential PNT services). Depending on the nature of the satellite targeted, the impacts on other state interests could be multiple and varied, encompassing effects on Earth and in space. Physical or non-physical harm might result from ASAT weapon use, including significant economic loss resulting from the denial of access to a satellite-provided service. A number of other legally protected rights might also be implicated, among them a state's right to neutrality.[71] Harm might also extend to rights and interests appertaining to all states. Significantly, the International Court of Justice (ICJ) has indicated that respect for the environment goes to assessing whether acts of self-defence conform to the requirements of necessity and proportionality,[72] with the UNGA underscoring that ASAT technologies might generally threaten the 'long-term sustainability of the outer space environment'.[73] Debris-creating ASAT weapon use could contaminate space in the long term and affect the ability of any and all states to operate in that domain and to benefit from the freedom to explore and use space peacefully, including placing satellites in orbit. The UNGA also reminds us that 'the

68 The ICJ has ruled that JAB proportionality may not exclude the use of nuclear weapons in the extreme circumstance of self-defence, where the very survival of a state would be at stake. *Nuclear Weapons* (n 26) paras 41–44, 97.

69 See Gardam (n 30) 17; O'Meara (n 33) 146–55. See also Nasu (n 10) 170–72.

70 Art IX obliges all states parties to the OST to conduct their space activities with due regard to the corresponding interests of all other states parties, including an obligation to avoid the harmful contamination of space.

71 See Wolff Heintschel von Heinegg, 'Neutrality and Outer Space' (2017) 93 Int'l L Stud 526; O'Meara (n 33) 147–53.

72 *Nuclear Weapons* (n 26) para 30.

73 UNGA Res 77/41 (n 4).

creation of long-lived orbital debris arising from the deliberate destruction of space systems increases the risk of in-orbit collisions and the potential for misunderstanding and miscalculations that could lead to conflict'.[74] This statement speaks to the wider possible impact of ASAT weapon use on international peace and security, in which all states have an interest. This peace and security is legally protected by Article 2(4) of the UN Charter and by strict adherence to the requirements of necessity and proportionality that limit states acting in self-defence. Consequently, ASAT weapon use that threatens this peace and security is likely to be regarded as disproportionate.

These conclusions have practical repercussions for states contemplating using ASAT weapons in self-defence. JAB proportionality obliges decision makers to consider the effects of the methods they use on Earth and in space. Given the risks associated with space debris, outside of extreme situations of self-defence threatening the existence of the state, JAB proportionality likely rules out using most, if not all, kinetic physical ASAT weapons. Non-kinetic alternatives, such as directed energy and electronic ASAT weapons, as well as cyber operations, might also be problematic for JAB proportionality compliance. This is so where the impacts on third-party interests are comparable to kinetic physical ASAT weapon use. An example is a cyber operation that causes a satellite to lose control and collide with another satellite or space object, resulting in damage and debris. Cyber operations might also impair the functionality of satellites providing essential PNT services either temporarily or permanently. Unless other satellites in a constellation ensure the continuation of the service, catastrophic consequences might ensue, potentially including loss of life. More generally, cyber operations raise particular concerns because attacks on a specific system may have repercussions for other systems and cause indiscriminate effects due to the interconnected nature of cyberspace.[75] Likewise, targeting mixed-use and multi-owner/user satellites with non-kinetic weapons might still result in potentially unquantifiable harm to states other than the aggressor, as well as to other non-state entities. With each example, affected third-party interests will weigh heavily on determinations of proportionality.

Not all potential targets will be multi-owner/user or mixed-use satellites, however, or will obviously implicate third-party interests. Where ASAT weapons do not cause space debris, these types of satellites might comprise less risky targets for defensive military operations, provided the effects of the targeting are contained. Yet, given the dangers of targeting satellites using any type of ASAT weapon, where self-defence can be effectively achieved by striking a target on Earth that has a nexus with the armed attack, rather than targeting a satellite in space, JAB proportionality logically requires that the former target be preferred over the latter. If a satellite can be neutralized by an

74 UNGA Res 75/36 (16 December 2020) UN Doc A/RES/75/36.

75 ICRC, 'International Humanitarian Law and Cyber Operations During Armed Conflicts' (28 November 2019) 2, 5, 6, 7. Raised in the context of applying IHL to the cyber domain but pertaining generally to cyber operations against mixed-use and multi-user satellites.

attack against a ground-based control node in a remote area, rather than targeting the satellite directly, this option must be taken.[76] Generally, targeting Earth-based objects avoids the risk of escalation in space and the associated threat to third parties and to international peace and security that JAB proportionality seeks to avoid.

5. CONCLUSION

Space is the 'province of all mankind'.[77] All states must be free to explore and use space peacefully, to benefit from satellites placed in space, and to have this communal resource protected from excessive military activities. Impingements on this freedom must be strictly controlled. Absent a multilateral ASAT weapons control treaty, the JAB (alongside IHL) must be regarded as an essential part of the international law framework limiting their use. Although states continue to develop new counterspace weapons and space is an ever-contested military domain, adherence to the requirements of JAB necessity and proportionality has the potential to limit ASAT weapon deployment. A clearer understanding of these JAB requirements, therefore, directly addresses pressing international concerns regarding the weaponization of space and the fear of wars between states in that domain. Given the unique nature of the space environment and the importance of satellites to the functioning of states and human societies, JAB compliance means that ASAT weapon use in self-defence is heavily restricted and may even be denied in all but the most extreme circumstances. Generally, targets of self-defence should be confined to the Earth. Where satellites are targeted with any form of ASAT weapon, states must take extreme caution. In addition to IHL, the JAB requires that methods employed to neutralize satellites be strictly controlled, limited to achieving a legitimate defensive purpose while minimizing harm to civilians and to third-party interests. Alternatives to ASAT weapons that cause physical damage should be preferred to avoid space debris. Weapons that only temporarily destabilize satellites or render them dysfunctional, or which are limited to interfering with or falsifying the transmission of signals to and from satellites, are most likely to comply with the JAB requirements. The same is true for targeted and limited cyber attacks on satellite-related computer networks. In each case, compliance depends on the harm caused by such 'soft kill' techniques to civilians and third-party interests.[78] Ultimately, adherence to JAB necessity and proportionality helps to avoid conflict and the escalation of conflict in space. JAB compliance underpins international peace and security and state aspirations of safeguarding space for peaceful purposes and ensuring its valuable resources continue to benefit all mankind.

76 Schmitt makes this point in respect of target selection and IHL precautions in attack requirements, but this conclusion arguably also applies to the requirements of JAB proportionality. Schmitt (n 10) 121.

77 OST (n 17) art I.

78 Methods at the lower end of the spectrum of military activity might not constitute uses of force and, therefore, do not require justification by reference to the JAB. However, they might constitute internationally wrongful acts. See UNGA, Articles on Responsibility of States for Internationally Wrongful Acts, annexed to UNGA, Res 56/83, UN Doc A/RES/56/83 (28 January 2002).

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

From Space Debris to Space Weaponry: A Legal Examination of Space Debris as a Weapon

Anna Blechová
PhD Student
Institute of Law and Technology
Faculty of Law
Masaryk University
Brno, Czech Republic
anna.blechova@law.muni.cz

Jakub Harašta
Assistant Professor
Institute of Law and Technology
Faculty of Law
Masaryk University
Brno, Czech Republic
jakub.harasta@law.muni.cz

František Kasl
Researcher
Institute of Law and Technology
Faculty of Law
Masaryk University
Brno, Czech Republic
frantisek.kasl@muni.cz

Abstract: Outer space represents an emerging and rapidly evolving domain that, until now, has remained free from armed conflicts. However, the stark reality is that the prospect of an arms race in space is no longer confined to dystopian imagination. This change in the environmental factual circumstances is substantiated by NATO's formal recognition of space as an operational domain and the renewed calls for a Treaty to Prevent an Arms Race in Space. Additionally, the symbolic characterization of space as the 'final frontier', as numerous scholars have described it, highlights the urgent need for analysis of the potential weaponization of the outer space domain.

Various potential weapons could be deployed in outer space. However, the primary objective of this paper is to investigate whether space debris in the highly commercialized and overpopulated Low Earth Orbit (LEO) could be weaponized. The example we focus on relates to the Kessler Syndrome and space debris generated by the use of (cyber) weapons.

The central message of this paper is that the potential triggering of Kessler Syndrome by creating space debris using space weapons should be considered an internationally wrongful act. Therefore, it should be taken into account during weapons reviews under Article 36 of Additional Protocol I to the Geneva Conventions (AP I). Moreover, the paper concludes that, in the context of the examined scenario and the rise of private entities within the space sector, it is also necessary to re-evaluate the legal framework regarding liability and responsibility in outer space as it relates to the triggering of the Kessler Syndrome.

Keywords: *space debris, weapon, outer space, liability, responsibility, Article 36 of Additional Protocol I*

1. INTRODUCTION

Outer space is garnering increasing attention as an operational domain. While the inherently curious nature of humanity is sometimes claimed to be the main driver for this, it is not the only one. Projection of power and spin-off technologies played a role in the past. Today, the increasing opportunities for commercial exploitation are also a factor. Space has become indispensable to the modern way of life.

Remarkably, since humanity's first venture into outer space in 1961, no active armed conflicts have occurred in this domain. The emphasis on peaceful activities in outer space within the Outer Space Treaty (OST) – the primary international law concerning space – has stood for almost 60 years.[1] However, terrestrial conflicts can extend into space, and the supposedly tranquil nature of the domain is slowly changing.[2]

Satellite infrastructure played a pivotal role in the early stages of Russia's blatantly unlawful full-scale invasion of Ukraine, manifesting in the cyber attack against the KA-SAT network in February 2022, which negatively impacted the situational awareness and communication capacities of the Ukrainian army and, thus, conferred an advantage to the Russian Federation.[3] Successful tests of kinetic anti-satellite

1 The OST has been in force since 1967. See Christopher Daniel Johnson, 'The Outer Space Treaty' *Oxford Research Encyclopedia of Planetary Science* (2018) <https://oxfordre.com/planetaryscience/display/10.1093/acrefore/9780190647926.001.0001/acrefore-9780190647926-e-43> accessed 13 March 2024.

2 Steven Freeland, 'The Peaceful Use of Outer Space: Protecting Life on Earth' (2023) Digital War <https://doi.org/10.1057/s42984-023-00065-w> accessed 13 March 2024.

3 'KA-SAT Network Cyber Attack Overview' (*Viasat*) <www.viasat.com/about/newsroom/blog/ka-sat-network-cyber-attack-overview/> accessed 10 May 2022; 'Case Study: Viasat Attack' (*Cyber Peace Institute*) <https://cyberconflicts.cyberpeaceinstitute.org/law-and-policy/cases/viasat> accessed 31 May 2023.

weapons (ASATs) carried out by China,[4] India,[5] and Russia[6] over the past few years are further demonstrations of the abovementioned changing dynamic.

Further evidence of the transformation discussed here can be found in official statements, international legal documents, and national space defence strategies that reflect these developments. Notably, the international community increasingly engages in discussions on outer space security and its implications in other areas. The United Nations General Assembly's First Committee on Disarmament and International Security adopted a resolution in November 2021 titled 'Reducing space threats through norms, rules and principles of responsible behaviours (L.52)'. This resolution advocates for a behaviour-based approach to addressing space security concerns and has led to the establishment of an Open-Ended Working Group (OEWG) to examine this issue. NATO has declared space an operational domain.[7] Moreover, various nations, including the US,[8] Germany,[9] and France,[10] have released space defence strategies declaring their interests in outer space.

Another factor to be taken into consideration is the increase in the pool of actors. Outer space is no longer the sole domain of nation-states. With the dwindling of funding for space programmes, the private sector has started to play an important role within the sector. This shift, accompanied by the proliferation of dual-use systems and an increasing number of commercial space objects in orbit, should be a driving force behind any discussion about the insufficiency of the existing legal framework.

The increasing traffic and a growing array of actors in outer space, coupled with the emergence of ASATs and cyber ASATs, necessitate a thorough examination of the

4 Brian Weeden, *2007 Chinese Anti-Satellite Test Fact Sheet* (Secure World Foundation 2010) <https://swfound.org/media/9550/chinese_asat_fact_sheet_updated_2012.pdf>; Shirley Kan, 'China's Anti-Satellite Weapon Test' (2007) CRS Report for Congress Order Code RS22652 <https://apps.dtic.mil/sti/pdfs/ADA468025.pdf>.

5 Ashley J Tellis, 'India's ASAT Test: An Incomplete Success' (*Carnegie Endowment for International Peace*) <https://carnegieendowment.org/2019/04/15/india-s-asat-test-incomplete-success-pub-78884> accessed 12 January 2024; 'India's Anti-Satellite Missile Test Is a Big Deal. Here's Why' (*Space.com*) <www.space.com/india-anti-satellite-test-significance.html> accessed 12 January 2024.

6 'Russian Direct-Ascent Anti-Satellite Missile Test Creates Significant, Long-Lasting Space Debris' (*U.S. Space Command*) <www.spacecom.mil/Newsroom/News/Article-Display/Article/2842957/russian-direct-ascent-anti-satellite-missile-test-creates-significant-long-last/> accessed 12 January 2024; 'Russia's Anti-Satellite Weapons: An Asymmetric Response to U.S. Aerospace Superiority' (*Arms Control Association*) <www.armscontrol.org/act/2022-03/features/russias-anti-satellite-weapons-asymmetric-response-us-aerospace-superiority> accessed 12 January 2024; 'Russia's Anti-Satellite Test Should Lead to a Multilateral Ban' (*SIPRI*) <www.sipri.org/commentary/essay/2021/russias-anti-satellite-test-should-lead-multilateral-ban> accessed 24 May 2022.

7 'Foreign Ministers Take Decisions to Adapt NATO, Recognize Space as an Operational Domain' (*NATO*) <www.nato.int/cps/en/natohq/news_171028.htm> accessed 12 January 2024.

8 '2020 Defense Space Strategy Summary' (*U.S. Department of State*) <https://media.defense.gov/2020/Jun/17/2002317391/-1/-1/1/2020_DEFENSE_SPACE_STRATEGY_SUMMARY.PDF>.

9 'The German Federal Government's Space Strategy' (*BMWK – Federal Ministry for Economic Affairs and Climate Action*) <www.bmwk.de/Redaktion/EN/Publikationen/Technologie/the-german-federal-governments-space-strategy.html> accessed 12 January 2024.

10 'Space Defence Strategy' (*Gouvernement*) <www.gouvernement.fr/sites/default/files/locale/piece-jointe/2020/08/france_-_space_defence_strategy_2019.pdf>.

associated legal ramifications. Our focus here is on an exploratory legal analysis of activities contributing to the escalating volume of space debris, which poses the risk of triggering the Kessler Syndrome.

The paper aims to investigate the evolving dynamics of outer space – particularly concerning space debris from (cyber) ASATs – and the potential triggering and damaging impact of the cascading effect of space debris known as Kessler Syndrome.[11]

Firstly, in Section 2, the paper introduces the requisite terminology and relevant legal frameworks. Secondly, Section 3 provides an in-depth examination of the area-denial capabilities of space debris. This inquiry is motivated by the realization that the repercussions of the proliferation of space debris, including that stemming from ASATs, have the potential to render outer space inaccessible to humanity – a circumstance incongruent with the established principles of international law. More specifically, the discussion focuses on the applicability of Article 36 of the AP I alongside the liability and responsibility regimes delineated within the framework of international law.

Building upon the discussion in this introduction, Section 4 provides an analysis of the different possible impacts of (cyber) weapons as regards the Kessler Syndrome and potential infringements upon international legal norms. The paper's central conclusion is that the existing scholarly literature pays insufficient attention to the Kessler Syndrome and its legal ramifications as they pertain to space-based weaponry. The paper's overarching aim is to offer new insights by not only recognizing the Kessler Syndrome as an undesirable outcome, a point well-documented in the existing literature, but also shedding light on its legal implications – a dimension that we contend has been underexplored.

2. SPACE DEBRIS

As humans, we are always concerned about what works and how it should work. We seem to be less concerned with things that used to work but are no longer needed. This is not meant as a critique; it is simply a consequence of the fact that in a modern and fast world, and especially today's world, we do not even have the capacity to be concerned. Space debris – meaning parts of old spacecraft, inoperable satellites, and human-made rubbish[12] but also natural objects such as meteoroids[13] – generally has no useful purpose. However, it is increasingly becoming a problem.

11 The Kessler Syndrome is discussed in detail in Section 2.

12 'Mission Monday: Five Fast Facts about the First American Spacewalk' (*Space Center Houston*) <https://spacecenter.org/mission-monday-five-fast-facts-about-the-first-american-spacewalk/> accessed 12 January 2024.

13 Linda Dawson, 'Space Debris as a Weapon' in Linda Dawson (ed), *War in Space: The Science and Technology Behind Our Next Theater of Conflict* (Springer International Publishing 2018) 46 <https://doi.org/10.1007/978-3-319-93052-7_4> accessed 12 January 2024.

The Space Debris Mitigation Guidelines of the Committee on the Peaceful Uses of Outer Space defines space debris as 'all man-made objects, including fragments and elements thereof, in Earth orbit or re-entering the atmosphere, that are non-functional'.[14] NASA uses a similar definition, which states that 'orbital debris is the term for any object in Earth orbit that no longer serves a useful function. These objects include non-operational spacecraft, derelict launch vehicle stages, mission-related debris, and fragmentation debris.'[15] Throughout this paper, we will use NASA's definition because we believe that even functional but unused objects should be considered space debris.

As of December 2023, the European Space Agency (ESA) has identified 36,500 space debris objects larger than 10 cm, 1 million objects ranging from 1 cm to 10 cm, and a staggering 130 million objects measuring between 1 mm and 1 cm.[16] While larger objects are more concerning, this does not mean that small objects are not a threat. Due to their high velocities (reaching speeds of 17,500 kph at a minimum), even the tiniest pieces of debris possess enough kinetic energy to disrupt critical systems on satellites or spacecraft. Notably, the International Space Station (ISS) suffered damage from 'a paint flake or small metal fragment no bigger than a few thousandths of a millimetre'.[17] Such an occurrence underscores the severity of encounters with even minor debris.[18] Whether big or small, space debris and space debris fragments are a threat to our presence in space that could, especially, grow in significance in relation to global navigation services, telecommunication services, weather forecasting, or climate change research.

A. Kessler Syndrome

The foremost concern linked to the proliferation of space debris is known as the Kessler Syndrome.[19]

14 United Nations Office for Outers Space Affairs, *Space Debris Mitigation Guidelines of the Committee on the Peaceful Uses of Outer Space* (United Nations 2010) <www.unoosa.org/res/oosadoc/data/documents/2010/stspace/stspace49_0_html/st_space_49E.pdf>.

15 Nicholas L Johnson, 'Orbital Debris Management & Risk Mitigation' (NASA) 6 <www.nasa.gov/wp-content/uploads/2018/12/692076main_orbital_debris_management_and_risk_mitigation.pdf>.

16 'Space Debris by the Numbers' (*European Space Agency*) <www.esa.int/Space_Safety/Space_Debris/Space_debris_by_the_numbers> accessed 12 January 2024.

17 'Impact Chip' (*ESA*, 12 May 2016) <www.esa.int/ESA_Multimedia/Images/2016/05/Impact_chip> accessed 12 January 2024.

18 ibid; Lizzie Plaugic, 'This Is What Happens When a Tiny Piece of Flying Space Debris Hits the ISS' (*The Verge*, 12 May 2016) <www.theverge.com/2016/5/12/11664668/iss-window-chip-space-debris-tim-peake> accessed 12 January 2024; Ted Muelhaupt, 'Space Debris and The Aerospace Corporation' (2015) 16 Crosslink 2–3.

19 Vilius Petkauskas, 'Why Hackers Destroying One Starlink Satellite Could Cause Orbital Armageddon' (*Cybernews*, 27 June 2022) <https://cybernews.com/editorial/why-hackers-destroying-one-starlink-satellite-could-cause-orbital-armageddon/> accessed 12 January 2024; Mike Wall, 'Kessler Syndrome and the Space Debris Problem' (*Space.com*, 15 November 2021) <www.space.com/kessler-syndrome-space-debris> accessed 12 January 2024; Francis Lyall and Paul Larsen, *Space Law: A Treaties* (2nd edn, Routledge 2018) 271.

As space debris collides with other objects, a cascading effect can be triggered that perpetuates the generation of additional space debris. This cascade effect, which is often likened to a domino effect or chain reaction, constitutes a pivotal aspect of the space debris issue. In theory, there is a potential scenario where the accumulation of space debris in our orbital vicinity reaches a critical mass, hindering the spacecraft's access to space. Compounding the issue is that the utility of satellites would also become severely constrained, thus imperilling their operational capabilities.[20]

This grim possibility is well illustrated by two notable events: The first was the People's Republic of China's ASAT test on Fengyun-1C in 2007. The test created more than 2,600 pieces of space debris over 10 cm across (over 100,000 in total), significantly accelerating the proliferation of space debris and, thus, further escalating the issue of orbital congestion.[21] The second was the 2009 collision between a U.S. satellite, Iridium 33, and a Russian satellite, Cosmos 2251,[22] which resulted in the creation of over 2,000 sizable fragments accompanied by numerous untraceable ones. Alarmingly, even three years post-collision, more than 90% of these fragments persisted in orbit, amplifying the hazards posed by an increasingly cluttered space environment.[23]

These events serve as harbingers, vividly illustrating the tangible consequences and potential ramifications of collisions or the use of weapons within the orbital realm.

As a worst-case scenario, the Kessler Syndrome could result in outer space becoming inaccessible to exploration and the cessation of vital space-related services, such as the Internet, telecommunications, or the global navigation satellite system (GNSS). Such an impact would effectively dismantle the contemporary world. The repercussions would extend beyond the loss of GPS navigation, impacting essential systems such as banking, and rendering the Internet unusable due to the absence of accurate time-stamping.[24]

20 Donald J Kessler and Burton G Cour-Palais, 'Collision Frequency of Artificial Satellites: The Creation of a Debris Belt' (1978) 83 Journal of Geophysical Research: Space Physics 2637; Jakub Drmola and Tomas Hubik, 'Kessler Syndrome: System Dynamics Model' (2018) 44–45 Space Policy 29; Donald J Kessler and others, 'The Kessler Syndrome: Implications to Future Space Operations' (*Semantic Scholar*, 2010) <www.semanticscholar.org/paper/THE-KESSLER-SYNDROME%3A-IMPLICATIONS-TO-FUTURE-SPACE-Kessler-Johnson/227655e022441d1379dfdc395173ed2e776d54ee> accessed 12 January 2024.

21 Leonard David, 'China's Anti-Satellite Test: Worrisome Debris Cloud Circles Earth' (*Space.com*, 2 February 2007) <www.space.com/3415-china-anti-satellite-test-worrisome-debris-cloud-circles-earth.html> accessed 12 January 2024.

22 RL Wang and others, 'Thinking Problems of the Present Collision Warning Work by Analyzing the Intersection between Cosmos 2251 and Iridium 33' *Proceedings of the 6th European Conference on Space Debris* (2013) <https://conference.sdo.esoc.esa.int/proceedings/sdc6/paper/45/SDC6-paper45.pdf>; Weeden (n 4).

23 Leonard David, 'Effects of Worst Satellite Breakups in History Still Felt Today' (*Space.com*, 28 January 2013) <www.space.com/19450-space-junk-worst-events-anniversaries.html> accessed 12 January 2024.

24 Charlotte Van Camp and Walter Peeters, 'A World without Satellite Data as a Result of a Global Cyber-Attack' (2022) 59 Space Policy 101458, 1–7.

It is worth noting that the Kessler Syndrome is not a rapid event akin to a 'space gunshot'. Rather, it resembles the slow congestion of the space environment with what is essentially rubbish. Moreover, according to Doboš and Pražák, the low-intensity Kessler Syndrome threshold, which requires significant planning effort when deploying new spacecraft and satellites, has already been reached.[25]

B. Space Debris and International Law

Space law applicable to space debris is mainly constituted by the Outer Space Treaty (OST) and the Liability Convention (LC).

The cornerstone of the OST is Article IX, which delineates the principle of non-contamination of space and establishes guidelines for conducting outer space activities while duly considering the legitimate interests of other states.[26] This article urges responsible behaviour in space endeavours to safeguard against contamination and to ensure that activities in outer space align with the mutual interests of all participating states. Moreover, based on Article IX and customary law on state responsibility,[27] we conclude that states are responsible for the space debris they produce because of the due regard principle.[28] With regard to this, it is crucial to emphasize that a space object and any space debris that originates from it share jurisdiction, carrying with it the corresponding legal implications. For instance, if a space object is registered in State A, the state assumes liability and responsibility not only for the object itself but also for any space debris that may originate from it.

According to the OST, if space debris is repurposed to intentionally cause damage or harm by State B, the liability and responsibility still lies with the state that initially launched the space object from which the debris originated. Such a principle underscores the accountability of the originating state for any subsequent misuse of space debris, emphasizing the need for careful and responsible conduct of outer space activities.

Within the framework of the LC, a pivotal question arises as to whether space debris would be classified as a space object, thereby falling under the purview of the LC. Article I of the LC defines a 'space object' as including not only the primary object but also its constituent components and associated launch vehicles. Space debris that could be considered a component of launch vehicles would arguably be subsumed under this definition.[29]

25 Bohumil Doboš and Jakub Pražák, 'Master Spoiler: A Strategic Value of Kessler Syndrome' (2022) 22 Defence Studies 123, 123.

26 John S Goehring, 'Can We Address Orbital Debris with the International Law We Already Have? An Examination of Treaty Interpretation and the Due Regard Principle' (2020) 85 Journal of Air Law and Commerce 309, 310–312.

27 For example, ARSIWA or the International Law Commission's articles on state responsibility: introduction, text, and commentaries.

28 Goehring (n 26) 336–337.

29 Isabella Henrietta Philepina Diederiks-Verschoor and Vladimír Kopal, *An Introduction to Space Law* (3rd rev edn, Kluwer Law International 2008) 128.

The classification of space debris as space objects can be approached from two distinct academic positions: spatialist and functionalist. The former approach hinges on location, designating anything beyond the airspace boundary as a space object. Conversely, the latter conceives space objects as operational space instruments regardless of their location. Thus, on the functionalist view, space debris may not be considered a space object precisely because it has lost its functionality.[30] While the spatialist theory seems more intuitive, the absence of an international consensus weakens its validity.

The absence of any mention of space debris in the LC or the OST can be attributed to the historical context of these documents.[31] At the time of their inception, space debris was not a prevalent concern, as space exploration was still in its nascent stages. Pelton asserts that the categorization of space debris as space objects is not straightforward and may require an amendment to the LC.[32]

The tranquil environment of outer space is undergoing changes, necessitating consideration of whether International Humanitarian Law (IHL) is applicable in outer space, which is defined as a peaceful domain.[33, 34] The applicability of these rules derives from Article III of the OST, which states that international law applies to the use of outer space. The International Committee of the Red Cross (ICRC) emphasizes the relevance of Article 2 of the Geneva Conventions, asserting that it is applicable to any conflict.[35]

In summary, although IHL and general space law (including the OST and the LC) represent distinct legal codes, they are not in conflict per se. With regard to legal stability and continuity, if an armed attack were to occur in space, IHL would not terminate or suspend general space law but complement it.[36] Regarding Article 36 of the AP I and its status as IHL, Dienelt claims it also applies in peacetime.[37] For these reasons, we will delve further into both legal concepts.

30 Gordon Chung, 'Jurisdiction and Control Aspects of Space Debris Removal' in Annette Froehlich (ed), *Space Security and Legal Aspects of Active Debris Removal* (Springer International Publishing 2019) 34–36 <https://doi.org/10.1007/978-3-319-90338-5_3> accessed 12 January 2024.

31 Lyall and Larsen (n 19) 272.

32 Joseph N Pelton, 'Legal Challenges Related to Active Orbital Debris Removal' in Joseph N Pelton (ed), *New Solutions for the Space Debris Problem* (Springer International Publishing 2015) 73 <https://doi.org/10.1007/978-3-319-17151-7_6> accessed 12 January 2024.

33 Preamble OST, art IV OST.

34 Michael Schmitt and Sqn. Ldr. Kieran Tinkler, 'War in Space and International Humanitarian Law' (*Just Security*, 9 March 2020) <www.justsecurity.org/68906/war-in-space-how-international-humanitarian-law-might-apply/> accessed 12 January 2024.

35 *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts – Recommitting to Protection in Armed Conflict on the 70th Anniversary of the Geneva Conventions, ref. 4427*, pp 32–34.

36 Article 3 of the draft articles on the effects of armed conflicts on treaties, with commentaries.

37 Anne Dienelt, 'The Shadowy Existence of the Weapons Review and Its Impact on Disarmament' (2018) 36 Sicherheit und Frieden (S+F) / Security and Peace 126, 126.

3. THE AREA-DENIAL CAPACITY OF SPACE DEBRIS

Currently, the generation of space debris in orbit is unavoidable. However, the increase in its amount and size distribution can be caused both intentionally and unintentionally, even to the point of triggering the Kessler Syndrome. In the scenario of intentional increase, space debris could be used as area-denial weapons, effectively rendering the 'ultimate high ground' useless. The proliferation of space debris, thus, poses a tangible threat that could possibly be used for geopolitical leverage, encapsulating the multifaceted risks associated with debris accumulation in outer space.

Space debris primarily emanates from anthropogenic activities. Abandoned, non-functional, and deteriorating spacecraft and satellites are significant contributors to the amount of space debris. Moreover, the utilization of ASATs, when involving the interception of and collisions with operational satellites or other man-made objects, is a substantial generator of space debris. Testing of anti-satellite capacities (either kinetic or cyber) against existing space debris (such as largely compact defunct satellites) escalates the creation of space debris and poses a heightened risk in the orbital vicinity.

Cybersecurity considerations – or the lack of it – are an additional factor. The commercialized and arguably overcrowded LEO[38] hosts satellites and mega-constellations with notably subpar measures against cybersecurity threats.[39] In 2023, NIST published a relevant cybersecurity standard,[40] and IEEE initiated a working group aimed at creating one.[41] However, there is still a notable lack of internationally recognized soft laws as well as hard laws, leaving the ecosystem vulnerable to cyber attacks or the deployment of cyber weapons. Malfunctions via cyber means might escalate the generation of space debris and eventually trigger the Kessler Syndrome. In such an environment, malicious actors might deploy cyber ASATs as counterparts to kinetic ASATs. While kinetic ASATs are typically detectable within minutes of launch, the average detection time for a cyber data breach is approximately 200 days.[42]

38 Diederiks-Verschoor and Kopal (n 29) 21–22; Chitra Sethi, 'The Commercial Future of Low-Earth Orbit' (*Tech Briefs*, 1 August 2022) <www.techbriefs.com/component/content/article/46276-the-commercial-future-of-low-earth-orbit> accessed 12 January 2024; 'Low-Earth Orbits Are Getting Crowded' (*ESA*) <www.esa.int/ESA_Multimedia/Images/2022/04/Low-Earth_orbits_are_getting_crowded> accessed 12 January 2024.

39 'Satellites Are Rife with Basic Security Flaws' (*Wired*, 20 July 2023) <www.wired.com/story/satellites-basic-security-flaws/>; 'Cybersecurity Threats in Space: A Roadmap for Future Policy' (*Wilson Center*, 8 October 2020) <www.wilsoncenter.org/blog-post/cybersecurity-threats-space-roadmap-future-policy> accessed 13 January 2024.

40 Matthew Scholl and Theresa Suloway, 'Introduction to Cybersecurity for Commercial Satellite Operations (2nd Draft)' (National Institute of Standards and Technology 2022) NIST Internal or Interagency Report (NISTIR) 8270 (Draft) <https://csrc.nist.gov/publications/detail/nistir/8270/draft> accessed 29 May 2022.

41 'P3349 – Space System Cybersecurity Working Group – The Project' (*IEEE SA*) <https://sagroups.ieee.org/3349/the-project/> accessed 13 January 2024.

42 Brendan I Koerner, 'Inside the OPM Hack: The Cyberattack That Shocked the US Government' (*Wired*, 23 October 2016) <www.wired.com/2016/10/inside-cyberattack-shocked-us-government/> accessed 12 January 2024.

Advanced concealment methods and the already mentioned lack of cybersecurity standards create a vulnerable environment that provides opportunities to intentionally increase the amount of debris over time to deny important capabilities and cause chaos.

The deliberate or inadvertent initiation of the Kessler Syndrome, resulting from either the intentional or unintentional creation of an extensive volume of space debris – such as through a destructive directed cyber attack on multiple mega-constellations or numerous directed kinetic attacks on a group of satellites – constitutes an internationally wrongful act. The violation stems from the fact that such an attack would constitute a breach of obligations outlined in the OST – particularly Article I – contravening the principle of exploration and use of outer space for all. Additionally, Article IX of the OST establishes the principle of due regard, emphasizing that exploration and use of outer space should not lead to potentially harmful interference. Closing access to outer space by triggering the Kessler Syndrome would breach both provisions of the OST.

A. Liability and Responsibility Regarding the Kessler Syndrome

Two pivotal legal concepts demand attention when considering the matter of collisions between space objects and the Kessler Syndrome: responsibility and liability. It is important to underline that these concepts originate from distinct temporal and contextual backgrounds, thereby rendering their contemporary application challenging.[43]

Primarily, our attention will be directed towards the complexities pertaining to the concept of responsibility. Article VIII of the OST stipulates that no state has the authority to destroy objects outside of its jurisdiction unless an agreement for a justifiable cause is in place. The article embodies a customary law obligation and explicitly confers jurisdiction and control to the state of registration, which assumes liability for the space object. Notably, the legal provision does not set a time limit for sovereignty, effectively allowing the state of registration to retain it indefinitely.[44] Hence, if State A were to destroy a space object belonging to State B, it would constitute a breach of the OST. Consequently, if the act were attributable to State A, it would qualify as an internationally wrongful act for which State A would be held responsible. Furthermore, since the registration of space objects is non-transferable, even in the case of damage to a non-functional or inactive space object (like space debris), states could be held responsible if acting outside of their jurisdiction.

43 Frans G von der Dunk, 'Liability versus Responsibility in Space Law: Misconception or Misconstruction?' *Proceedings of the 34th Colloquium on the Law of Outer Space* (1992) 363–371 <https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1020&context=spacelaw>; Bartosz Ziemblicki and Yevgeniya Oralova, 'Private Entities in Outer Space Activities: Liability Regime Reconsidered' (2021) 56 Space Policy 101427.

44 Chung (n 30) 33–34.

Considering the challenging task of attributing cyber ASATs to the correct entities, as well as the responsibilities outlined in Article IV of the OST, helps us to recognize the possibility of a concerning scenario involving exploitation by malicious actors. In this scenario, owners of cyber ASATs successfully target a space object – potentially even space debris – belonging to State B, resulting in damage. In such a case, responsibility for the damage could fall upon State B (the affected state), the launching state, and the registrar of the targeted space object.

Furthermore, it is crucial to highlight that, under the current framework established in Article VI of the OST, states are responsible for the actions of both state and non-state actors. In light of the ongoing commercialization of outer space, this presents a significant legal concern. For instance, if a commercial entity that launched and registered its space object in State A triggers the Kessler Syndrome through testing a cyber ASAT on its own satellite, State A would bear objective responsibility.

While some states might have national space legislation to address such scenarios, international sanctions for wrongful acts primarily target states. Consequently, even if a state has laws to hold private companies accountable, it might face challenges in transferring the responsibility (or even liability, see below) on the level of national law to the private entity.[45]

Having considered responsibility, we turn to the concept of liability. If it is the case that the LC applies to space debris within space law, then liability is governed by two distinct regimes. The first is the absolute liability principle, as outlined in Article II of the LC, which applies to any damage caused to the surface of the Earth or to aircraft in flight. The second regime is fault-based liability, as detailed in Article III of the LC, which addresses damage occurring elsewhere than on the Earth's surface, particularly damage to space objects.

When considering the possibility of an intentional triggering of the Kessler Syndrome with respect to damages to other satellites hit, it is conceivable that the breach of the OST could be established. That means that if such a breach were attributable, it would constitute an international wrongful act. However, in assessing liability regimes, it is crucial to evaluate the potential impact of the Kessler Syndrome on the Earth's surface. Despite its monumental repercussions on space exploration and potential adverse effects on terrestrial living standards, that absolute liability would be invoked seems improbable. In terms of the fault-based regime, proving fault for the state that launched cyber ASATs that triggered the Kessler Syndrome and caused damage to other space objects poses what appears to be an insurmountable challenge because of

45 See Bartosz Ziemblicki and Yevgeniya Oralova, 'Private Entities in Outer Space Activities: Liability Regime Reconsidered' (2021) 56 Space Policy 101427; Paul Dempsey, 'National Laws Governing Commercial Space Activities: Legislation, Regulation, & Enforcement' (2016) 36 Northwestern Journal of International Law & Business 1.

the low probability that a strong causal link could be established or sufficient evidence found.

B. Weapons Review under Article 36 of the AP I

Under Article 36 of the AP I, 'new weapons, methods, or means of warfare' must be reviewed with due regard to the prevention of the use of weapons 'that would violate international law in all circumstances and to impose restrictions on the use of weapons that would violate international law in some circumstances'.[46] ASATs and cyber ASATs, which have the potential to significantly increase the amount of space debris and arguably bring the environment closer to triggering the Kessler Syndrome, require weapons review while taking into account their broader (potentially mid- to long-term) impact.

While the terms used in the wording of Article 36 of the AP I are not precisely defined, a weapon typically implies an offensive capability applicable against a military object or enemy combatant.[47] Since ASAT and cyber ASATs could be applicable in this context, we contend that both fall within this scope. As there is no specialized treaty or customary law that specifically addresses ASATs and cyber ASATs by formulating rules for their deployment or use, the subsequent evaluation focuses on general rules applicable to all weaponry.

Since the deliberate or inadvertent creation of space debris might lead to extensive, enduring, and significant harm to the natural environment and interests of humanity in outer space,[48] we posit that ASATs and cyber ASATs might infringe upon the prohibition on damaging the environment outlined in Article 35(3) and Article 55 of the AP I. Furthermore, contaminating outer space violates Article IX of the OST, which appears inevitable when creating a significant amount of space debris. Thus, we are convinced that the possibility that such environmentally destructive effects could be triggered must be considered within any review under Article 36 of the AP I.

ASATs and cyber ASATs are not new weapons; nevertheless, we claim that the use of ASATs or cyber ASATs to trigger the Kessler Syndrome and intentionally create space debris could be classified as a new means and method of warfare because doing so might alter such weapons' capabilities and effects.[49]

46 'A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977: International Committee of the Red Cross Geneva, January 2006' (2006) 88 International Review of the Red Cross 931, 4.

47 Justin McClelland, 'The Review of Weapons in Accordance with Article 36 of Additional Protocol I' (2003) 85 International Review of the Red Cross 397, 404.

48 Chung (n 30) 36–37.

49 Vincent Boulanin and Maaike Verbruggen, 'Article 36 Reviews: Dealing with the Challenges Posed by Emerging Technologies' (Stockholm International Peace Research Institute 2017) 4 <www.sipri.org/sites/default/files/2017-12/article_36_report_1712.pdf>.

While the argumentation is straightforward for kinetic ASATs, there are exceptions for cyber ASATs. Solely disrupting the functionality of a space object may not lead to its destruction and the generation of a large amount of space debris.[50] The outcome of the weapons review hinges on the distinction between traditional ASATs, potentially illegal due to their destructive nature, and cyber ASATs, which are primarily focused on disruption and so might be considered legal, thus influencing their regulatory status. The Kessler Syndrome depends on the density of space debris – and although non-functional space objects are, by definition, considered space debris, their impact on the environment can differ. Seizing control of satellites by cyber means and adjusting their orbits to create collisions is problematic regarding triggering the Kessler Syndrome, whereas shutting them down seems to be a preferable option for society with regard to the lower mid-term and long-term impact on the outer space environment.

C. The (Legal) Issues of Space Debris Weaponization via Triggering the Kessler Syndrome

Alongside the legal challenges – such as liability, responsibility, and the weapons review under Article 36 of the AP I – there are additional formidable obstacles. The first pertains to identifying the origin of space debris. The second revolves around the complexities of legitimizing the targeting of space objects.

The primary challenge in legally categorizing space debris as a weapon lies in detecting its origin.[51, 52] While it may be feasible but very difficult to trace the satellite or space object from which the debris originated – whether through kinetic or cyber means – identifying the creator of the debris presents a distinct challenge.[53] With kinetic ASATs, it is plausible that the state or entity that launched the object could be identified and the object's trajectory tracked. However, in the case of cyber attacks or cyber ASATs, attributing the attack to the correct entity becomes exceedingly difficult. This echoes the complexities encountered in identifying perpetrators within cyber attacks on Earth. A parallel issue emerges in pinpointing the state responsible for triggering the Kessler Syndrome.

The question is whether it is legitimate to target space debris under IHL. Article 52(2) of the AP I delineates military objectives, defining them as objects that, due to their nature, location, purpose, or use, contribute effectively to military actions, offering a clear military advantage in their destruction, capture, or neutralization. Given that space objects are predominantly dual-use in nature, they might be perceived as

50 James Pavur and Ivan Martinovic, 'The Cyber-ASAT: On the Impact of Cyber Weapons in Outer Space' *11th International Conference on Cyber Conflict (CyCon)* (IEEE 2019) 5–7 <https://ieeexplore.ieee.org/document/8756904/> accessed 12 January 2024.

51 Alessandra Celletti, Giuseppe Pucacco and Tudor Vartolomei, 'Reconnecting Groups of Space Debris to Their Parent Body through Proper Elements' (2021) 11 Scientific Reports 1.

52 Pelton (n 32) 73–74.

53 Di Wu and Aaron J Rosengren, 'An Investigation on Space Debris of Unknown Origin Using Proper Elements and Neural Networks' (2023) 135 Celestial Mechanics and Dynamical Astronomy 44.

legitimate targets.[54] On the other hand, since any space debris object is, by definition, a non-functional or unused space object, establishing that the conditions outlined in Article 52(2) of the OST are met could prove challenging.

4. DISCUSSION

A. The Illegality of Triggering the Kessler Syndrome

The triggering of the Kessler Syndrome is undesirable since it could violate Article I and Article IX of the OST due to its environmental consequences. Consequently, we contend that the triggering of the Kessler Syndrome could be construed as a violation of international law that leads to an international wrongful act. If this is indeed the case, considering the possibility that the Kessler Syndrome may have already commenced,[55] it would be plausible to apply international responsibility as stated in the Draft Articles on Responsibility of States for Internationally Wrongful Acts. Therefore, hypothetically, when states such as China or Russia tested their ASATs, they likely violated international obligations and thus committed an international wrongful act.

States need to consider the potential classification of triggering the Kessler Syndrome as a wrongful act on various levels, particularly concerning the development of new weapons. It is a crucial consideration when determining the area-denial capacity of space debris (via the Kessler Syndrome), influencing whether a space weapon can undergo a weapons review under Article 36 of the AP I. Consequently, failure to conduct a weapons review could be deemed a violation of IHL. However, given that space is designated as a peaceful arena in the preamble and Article IV of the OST, the question of whether IHL is applicable in this domain is pertinent.[56] Nonetheless, contemporary perspectives hold that 'peaceful' does not equate to 'non-military'.[57] This corresponds to Article III of the OST, which stipulates that activities in outer space should be conducted in accordance with international law.[58] Thus, it follows

54 Almudena Azcárate Ortega, 'Not a Rose by Any Other Name: Dual-Use and Dual-Purpose Space Systems' (*Lawfare*, 5 June 2023) <www.lawfaremedia.org/article/not-a-rose-by-any-other-name-dual-use-and-dual-purpose-space-systems> accessed 12 January 2024; Jakub Pražák, 'Dual-Use Conundrum: Towards the Weaponization of Outer Space?' (2021) 187 Acta Astronautica 397.

55 Andrea Gini, 'Don Kessler on Envisat and the Kessler Syndrome' *Space Safety Magazine* (Winter 2012) <http://www.spacesafetymagazine.com/space-debris/kessler-syndrome/don-kessler-envisat-kessler-syndrome/> accessed 12 January 2024.

56 Jasani Bhupendra and Maria A Lunderius, 'Peaceful Uses of Outer Space-Legal Fiction and Military Reality' (1980) 11 Bulletin of Peace Proposals 57.

57 Tinkler (n 34).

58 Cassandra Steer and Dale Stephens, 'International Humanitarian Law and Its Application in Outer Space' in Cassandra Steer and Matthew Hersch (eds), *War and Peace in Outer Space: Law, Policy, and Ethics* (Oxford University Press 2020) 51–53 <https://doi.org/10.1093/oso/9780197548684.003.0002> accessed 12 January 2024.

that if the use of force occurs in outer space, IHL will apply in full, regardless of time and place.[59]

B. The Differences between ASATs and Cyber ASATs and Their Legal Implications

The militarization of outer space, once considered dystopian, is now a reality as spacefaring nations seek to deploy weapons beyond Earth's atmosphere.[60] A recent successful ASAT test, despite its achievements, underscores the significant hazards of generating extensive space debris and the problems with the lack of enforceable international law. Since ASATs are inherently destructive, so, too, they inherently contribute to the risk of triggering the Kessler Syndrome. In recognition of this concern, there is an ongoing discussion about the prohibition of such weapons led by the United States.[61]

An alternative worth exploring at this juncture is cyber ASATs. These possess unique advantages, particularly in terms of cost-effectiveness and the difficulty of attributing them to the correct entity.[62] From an environmental and legal standpoint, a key feature of cyber ASATs is their ability to 'turn off' satellites through various means, mitigating the risks of the creation of space debris.[63] However, cyber ASATs are not without their challenges – unintended impacts on multiple satellites within a mega-constellation and their potential availability to countries without established space capabilities are notable concerns. When considering cyber ASATs, it is also necessary to highlight that there is a greater chance that they would be owned or supported by non-state actors and private entities. This is an important difference compared to ASATs, which are mostly state-driven projects.

Given their destructive nature and the risks associated with triggering the Kessler Syndrome, ASATs could be deemed illegal. In contrast, cyber ASATs, which focus on disruption, might contribute less to the Kessler Syndrome and could possibly be considered legal. Hence, a shift in focus from the conventional arms space race to a cyber arms space race may be prudent.

59 *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts – Recommitting to Protection in Armed Conflict on the 70th Anniversary of the Geneva Conventions, ref. 4427*, pp 32–34.

60 David E Sanger and Julian E Barnes, 'U.S. Fears Russia Might Put a Nuclear Weapon in Space' *New York Times* (17 February 2024) <www.nytimes.com/2024/02/17/us/politics/russia-nuclear-weapon-space.html> accessed 10 March 2024.

61 Theresa Hitchens, 'Debris from ASAT Tests Creating "Bad Neighborhood" in Low Earth Orbit: Analyst' (*Breaking Defense*, 16 June 2023) <https://breakingdefense.sites.breakingmedia.com/2023/06/debris-from-asat-tests-creating-bad-neighborhood-in-low-earth-orbit-analyst/> accessed 12 January 2024; Jeff Foust, 'More Countries Encouraged to Commit to Halt Destructive ASAT Tests' (*SpaceNews*, 15 June 2023) <https://spacenews.com/more-countries-encouraged-to-commit-to-halt-destructive-asat-tests/> accessed 12 January 2024.

62 Pavur and Martinovic (n 50) 5–7.

63 ibid 6.

C. The Fault-Based Nature of Traffic in Outer Space and the Production of Space Debris

The surge in the number of private entities for which states bear responsibility and liability presents numerous challenges. The existing legal framework, grounded in Article VI of the OST, asserts that states are responsible for national activities conducted by governmental or non-governmental entities. However, as the intentions of states and private companies may be inherently incompatible and distinct, there is a growing call for a re-evaluation of the relevant legal framework.[64]

Concerning the generation of space debris and the potential triggering of the Kessler Syndrome, the approaches of states and private companies could not be more different. Consider a hypothetical scenario where a private entity competitively destroys a space object launched and registered in a different state, yet due to the absence of adequate regulation, it is not held liable or responsible. Even if the company's motive was not to trigger the Kessler Syndrome but to gain a competitive advantage, it could inadvertently do so. Under the current legal framework, the relevant states would be held accountable. This possibility, as well as analogous scenarios likely to emerge with the burgeoning space industry, necessitates a re-evaluation not only of liability in general but also in the light of the possibility of triggering the Kessler Syndrome.

5. CONCLUSION

The escalating cyber threat in outer space raises significant concerns, particularly regarding the potential deployment of ASATs, which increase the amount of space debris and may, thus, conceivably trigger the Kessler Syndrome.

The looming threat of the Kessler Syndrome provides a compelling argument for the prohibition of ASATs, as assessed through a weapons review under Article 36 of the AP I, and for greater attention to the broader issue of space weaponization. It is crucial to differentiate between traditional ASATs and cyber ASATs. While the former would likely fail weapons review due to the substantial environmental consequences of their use, the latter, which could have a comparatively smaller environmental impact, might pass such assessments. This distinction prompts a re-evaluation, shifting our focus from an arms race in outer space to a cyber arms race in the same domain.

Moreover, the examination of liability and responsibility with respect to the Kessler Syndrome reveals the challenges of attributing (cyber) ASAT attacks to the correct entities and holding states accountable. This paper underscores the imperative need

64 Alexander P Reinert, 'Updating the Liability Regime in Outer Space: Why Spacefaring Companies Should Be Internationally Liable for Their Space Objects' (2020) 62 William & Mary Law Review 325; Ziemblicki and Oralova (n 45); Biswanath Gupta and Raju KD, 'Understanding International Space Law and the Liability Mechanism for Commercial Outer Space Activities – Unravelling the Sources' (2019) 75 India Quarterly 555.

for an updated legal framework capable of addressing the evolving landscape of space activities, especially considering the increasing participation of private entities that have different intentions from state actors.

In summary, the creation of space debris and the potential triggering of the Kessler Syndrome through the use of weapons in space represents an unregulated issue that is insufficiently covered within the current literature. This discussion opens a dialogue on the existing legal framework, treating the Kessler Syndrome not merely as a remote possibility but as an undesirable condition that could effectively 'close outer space' for humankind. Importantly, each step towards such a state should be viewed not just as a negative factual development but also as an internationally wrongful act, emphasizing the need for proactive legal measures and international cooperation in preserving the peaceful use of outer space.

ACKNOWLEDGEMENTS

Our special thanks go to our families and relatives for their support, motivation, and toleration of our work schedules. We would also love to thank Jakub Vostoupal because he came when Anna called for aid.

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

Military Psychological Operations in the Digital Battlespace: A Practical Application of the Legal Framework

Anastasia Roberts
Independent Legal Consultant
United Kingdom

Adrian Venables
Senior Lecturer
Cyber Security Masters
Programme Manager
Tallinn University of Technology
Estonia

Abstract: This paper aims to clarify the legal framework for military psychological operations (PsyOps) in the digital battlespace during international armed conflict (IAC), with a particular focus on civilian protection. This is to support its practical application in military training and planning. The impetus for this paper is twofold: the increasing complexity and scope of PsyOps from a technological perspective and the resurgence of IAC, impacting military training and planning. To provide practical context, the North Atlantic Treaty Organization (NATO) and its member states are used as a focal point for this discussion. The roles of both international human rights law (IHRL) and international humanitarian law (IHL) in regulating military PsyOps are considered. The inherent tension between the two bodies of law, due to their different origins, is then discussed. This paper concludes that the two bodies of law can in fact be reconciled. This is by using IHL provisions that regulate PsyOps from a civilian protection perspective to encompass broader human rights concerns. This paper advocates for the development of a comprehensive assessment and authorization process for training and planning for military PsyOps in IAC based on IHL.

Keywords: *international armed conflict, international humanitarian law, international human rights law, civilian protection, psychological operations*

1. INTRODUCTION

This paper aims to clarify the legal framework for military psychological operations (PsyOps) in the digital battlespace during international armed conflict (IAC). This is to support its practical application in military training and planning. To provide context, the North Atlantic Treaty Organization (NATO) and its member states will be used as a focal point for this discussion.

There is no universally agreed definition of PsyOps. However, this paper will use the NATO definition: *'Planned activities using methods of communication and other means directed at approved audiences in order to influence perceptions, attitudes and behaviour, affecting the achievement of political and military objectives.'*[1]

The impetus for this paper is twofold. Firstly, the increasing complexity and scope of PsyOps from a technological perspective. Secondly, the resurgence of IAC, impacting military training and planning.

PsyOps are certainly not new, but advances in information and communications technologies (ICT) have expanded their potential reach and depth. In terms of reach, audiences selected for influence activity can now be accessed at scale through the internet, particularly with mobile technology facilitating virtually constant user access. In terms of depth, influence can be effected in more targeted and intrusive ways. For example, in the context of the Russia–Ukraine conflict, some Ukrainian military personnel reportedly received personalized messages on their own devices, coercing them to surrender.[2] Digitalization has also enabled more covert means of influence, to the extent that audiences may not even realize that they are being influenced. For example, in 2022, Meta (formerly Facebook) announced that it had removed a number of Russia-sponsored fake social media accounts. These were masquerading as credible news agencies and disseminating false reporting, or disinformation, about the situation in Ukraine.[3]

The Israeli response to the Hamas attacks of 7 October 2023 has also been marked by sophisticated PsyOps on both sides. Israel has employed its covert 'Influence Unit' to both shape the media's perception of the war and target Hamas terrorists.[4] Hamas

1 NATO Allied Joint Publication 3.10.1, *Allied Joint Doctrine for Psychological Operations with UK National Elements* (edn B, vers 1, 2014) para 0102 <https://www.gov.uk/government/publications/ajp-3101-allied-joint-doctrine-for-psychological-operations> accessed 4 March 2024.

2 Matthew Roscoe, 'Russia's Special Services Accused of Sending Threatening Messages to Ukrainian Soldiers' (*EuroWeekly News*, 8 June 2022) <https://euroweeklynews.com/2022/06/08/russia-threatening-message-ukrainian-soldiers/> accessed 4 March 2024.

3 Dan Milmo, 'Facebook Takes Down Ukraine Disinformation Network and Bans Russian-Backed Media' *The Guardian* (28 February 2022) <https://www.theguardian.com/technology/2022/feb/28/facebook-takes-down-disinformation-network-targeting-ukraine-meta-instagram> accessed 4 March 2024.

4 Eric Cortellessa and Vera Bergengruen, 'Inside the Israel-Hamas Information War' *Time* (22 December 2023) <https://time.com/6549544/israel-and-hamas-the-media-war/> accessed 4 March 2024.

has also employed PsyOps under the guise of hacktivist campaigns of #OpIsrael and #FreePalestine using a range of social media outlets.[5]

It could be argued that these technological advances do not fundamentally alter the underlying legal regime for PsyOps and that the legal issues raised by PsyOps have not changed. In the purest sense, this is true. However, the difference lies in the practical application of the legal regime to more complex means of influence. Leaflet drops and radio-in-a-box can generally be limited geographically and addressed to specific audiences, enabling their effects to be anticipated. The use of social media, with its potential for rebroadcasting and manipulation, is more difficult to control and assess.

The second stimulus for this paper is the fact that NATO and its member states are currently having to readjust their thinking, training and planning to fully encompass the potential for IAC. This is after decades of focusing on non-international armed conflict (NIAC) and situations below the threshold of armed conflict. The Russia–Ukraine conflict has demonstrated that IAC is still a reality. Estonia's Foreign Intelligence Service has only recently issued a warning that Russia is preparing for a war with NATO within the next decade.[6] The potential for fighting between Israel and Hamas to escalate to involve third-party states also supports the increased threat of IAC, as does the increasing tension between China and Taiwan.

NATO and its member states must start to consider how they will employ PsyOps in full-scale war fighting as opposed to the more limited context of counter-insurgency seen in Iraq and Afghanistan. This requires a comprehensive understanding of the legal regime governing PsyOps in IAC, which is different from that applicable in sub-threshold situations or even, to an extent, in NIAC.

Understanding the legal regimes applicable to different types of operations can be challenging for military personnel. This has not been helped by the fact that the line between peacetime and armed conflict, and their different legal regimes, has become increasingly blurred in common understanding in recent years, with constant references to 'grey-zone activity'. Furthermore, there is an ongoing tension between international humanitarian law (IHL) and international human rights law (IHRL), particularly in the context of IAC. IHL is the specialized body of law in armed conflict, but IHRL also applies as a matter of legal principle, subject to complex jurisdictional issues.

5 '#OpIsrael, #FreePalestine and #OpSaudiArabia – How Cyber-Threat Actors Coordinate PSYOPS Campaigns with Kinetic Military Actions' (*Resecurity*, 9 October 2023) <https://www.resecurity.com/blog/article/opisrael-freepalestine-and-opsaudiarabia-how-cyber-actors-capitalize-on-war-actions-via-psy-ops> accessed 4 March 2024.

6 Sergey Goryashko, Pierre Emmanuel Ngendakumana, 'Russia Gearing Up for Decade-Long Duel with West, Estonia Warns' *Politico* (13 February 2024) <https://www.politico.eu/article/russia-prepares-for-decade-long-confrontation-with-west-estonia-warns/> accessed 4 March 2024.

Specific human rights concerns have been raised about the impact of PsyOps on civilians in armed conflict, in terms of privacy[7] as well as freedom of opinion and expression.[8] In this context, IHL has been criticized for not sufficiently regulating PsyOps, addressing them only *'tenuously and non-systematically'*[9] and taking *'a remarkably lenient approach'*[10] to them. How should this tension between IHL and IHRL be approached in military training and planning for PsyOps in IAC? A number of national military manuals provide guidance on PsyOps, but not to this level of detail. Furthermore, many were written some time ago. The well-known lawyer's response that the answer depends on the circumstances is unhelpful.

This paper will attempt to clarify the legal regime for PsyOps in IAC for the purposes of military training and planning. Given the concerns raised about civilian protection, it will focus on this aspect of the regime. To provide practical context, it will first consider how NATO and its member states employ and regulate PsyOps from a doctrinal perspective. The paper will then consider IHRL concerns before moving on to consider whether IHL can in fact address these concerns, in an effort to reconcile the two bodies of law. This paper concludes that the two bodies of law can be reconciled and advocates for the development of a comprehensive assessment and authorization process for training and planning military PsyOps based on IHL.

To be clear, this paper does not seek to break fresh academic ground but merely to re-establish some order in what has become a confused space by proposing a pragmatic way forward in terms of military training and planning for PsyOps in IAC.

2. MILITARY DOCTRINE AND GUIDANCE

By examining available NATO and national doctrine and guidance, it is possible to extrapolate common approaches to the use and regulation of PsyOps, which are outlined below. NATO doctrine is particularly useful as it indicates a degree of consensus between a large body of states, albeit in broad and non-operationally specific terms.

7 Russell Buchan and Asaf Lubin (eds), *The Rights to Privacy and Data Protection in Times of Armed Conflict* (NATO CCDCOE Publications 2022) 3.

8 OHCHR 'Disinformation and Freedom of Opinion and Expression During Armed Conflicts: Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (12 August 2022) UN Doc A/77/288 <https://www.ohchr.org/en/documents/thematic-reports/a77288-disinformation-and-freedom-opinion-and-expression-during-armed> accessed 4 March 2024.

9 Robin Geiss and Henning Lahmann, 'Protecting the Global Information Space in Times of Armed Conflict' (2021) The Geneva Academy of International Humanitarian Law and Human Rights Working Paper 8 <https://www.geneva-academy.ch/news/detail/ 452-three-papers-map-contentious-issues-related-to-the-application-of-international-law-to-military-cyber-operations> accessed 4 March 2024.

10 Eian Katz, 'Liar's War: Protecting Civilians from Disinformation during Armed Conflict' (2020) 102 [914] IRRC 659, 663.

A. Audiences

Those selected for influence effect are known as target or targeted audiences (TAs). TAs may include civilians as well as adversary audiences. NATO PsyOps doctrine refers to TAs as ranging *'from populations to decision-makers at all levels'*.[11] The US Law of War Manual[12] supports this broad TA base, specifically referring to civilian and neutral audiences, as does the French Law of Military Operations Manual.[13]

B. Analysis

TAs are analysed to understand how best to influence them. This is known as target audience analysis (TAA), defined in NATO doctrine as *'the focused examination of targeted audiences to create desired effects'*.[14] PsyOps relies on *'extensive information and intelligence'* about the TAs. This includes their location, vulnerabilities, susceptibilities, strengths and weaknesses, and social and cultural characteristics, among other requirements.[15]

C. Enablement

Cyber operations may be used as an enabler for PsyOps activity. NATO cyber doctrine explains that cyber operations can support PsyOps by *'providing both a vector for deploying information and effects that influence targeted audiences'*.[16] The French manual refers to the use of messaging through internet-based social networks as a means of effecting influence.[17]

D. Targeting

NATO doctrine is clear that PsyOps capability does not sit in isolation but also contributes to other military activities.[18] One of these is targeting, which is the employment of lethal or non-lethal capability against selected adversary targets to create specific physical, virtual or cognitive effects.[19] In NATO targeting doctrine, the definition of target includes a person or group of people, including their mindset, thought processes, attitudes and behaviours.[20] More recent doctrine specifically refers

11 NATO (n 1) para 0126.

12 US Department of Defense, *Law of War Manual* (12 June 2015, updated July 2023) para 5.26.1.2 <https://media.defense.gov/2023/Jul/31/2003271432/-1/-1/0/DOD-LAW-OF-WAR-MANUAL-JUNE-2015-UPDATED-JULY%202023.PDF> accessed 4 March 2024.

13 Le ministère des Armées, *Manuel de Droit des Opérations Militaires* (2022) para 8.1.3.1. <https://tinyurl.com/2an95hzs> accessed 4 March 2024.

14 NATO Allied Joint Publication 10.1, *Allied Joint Doctrine for Information Operations with UK National Elements* (edn A, vers 1, 2023) LEX-11 <https://www.gov.uk/government/publications/allied-joint-doctrine-for-information-operations-ajp-101> accessed 4 March 2024.

15 NATO (n 1) para 0118.

16 NATO Allied Joint Publication 3.20, *Allied Joint Doctrine for Cyberspace Operations* (edn A, vers 1, 2020) para 1.32 <https://www.gov.uk/government/publications/allied-joint-doctrine-for-cyberspace-operations-ajp-320> accessed 4 March 2024.

17 Le ministère des Armées (n 13) para 8.1.3.1.

18 NATO (n 1) paras 0124–0135.

19 NATO Allied Joint Publication 3.9, *Allied Joint Doctrine for Joint Targeting*, (edn B, vers 1, 2021) para 1.2.2 <https://www.gov.uk/government/publications/allied-joint-doctrine-for-joint-targeting-ajp-39a> accessed 4 March 2024.

20 ibid LEX-17.

to audiences or organizations in the target definition.[21] PsyOps capability supports target analysis and advises on the most effective means of creating the desired effects as well as directly delivering influence effect when this is the selected means.

E. Attribution

PsyOps have traditionally been categorized according to their attributability; that is to say whether they are ascribable to a source. White PsyOps involve fully attributed products. Grey PsyOps involve products that do not specifically reveal their source. Black PsyOps involve products that appear to emanate from a source other than the true one. NATO PsyOps doctrine states that PsyOps are generally attributable to NATO, to preserve credibility. However, from a national perspective, the UK position is that UK PsyOps are *'predominantly, but not exclusively, "white"'*.[22]

F. Truthfulness

NATO's position is that PsyOps products must be based on true information and that using false information is counter-productive to the long-term credibility and success of PsyOps.[23] In contrast, the German Law of Armed Conflict Manual appears to acknowledge the reality that sometimes PsyOps may not be truthful. It states that *'it is permissible to exert political and military influence by spreading – even false – information to undermine the adversary's will to resist and to influence their military discipline (e.g. calling on them to defect, surrender or mutiny)'*.[24]

G. Authorization

In accordance with NATO PsyOps doctrine, TAs and PsyOps effects must be approved by the North Atlantic Council through the submission of an operational plan.[25] This will include any rules of engagement for PsyOps activity.[26] The operational plan will provide the overarching regulatory framework for PsyOps and detail its interaction with other capabilities and functions, including the targeting process. In the context of the operation itself, specific PsyOps plans and products must be approved at the appropriate level of command, including any enabling cyber operations.[27] Approval is subject to consideration of any predicted cognitive, virtual and physical impact.[28] NATO doctrine also mandates the close involvement of a legal adviser in both lethal and non-lethal targeting, including PsyOps, to ensure compliance with the legal framework.[29]

21 NATO (n 14) LEX-11.
22 NATO (n 1) para 0115.
23 NATO (n 1) para 0114.
24 German Joint Service Regulation (ZDv)15/2, *Law of Armed Conflict Manual* (1 May 2013) para 487 <https://www.onlinelibrary.iihl.org/wp-content/uploads/2021/05/GER-Manual-Law-of-Armed-Conflict.pdf> accessed 4 March 2024.
25 NATO (n 1) para 0509.
26 NATO (n 1) para 0311.
27 ibid.
28 NATO (n 14) para 4.35.
29 NATO (n 19) para 1.6.1; NATO (n 14) para 3.9.

3. IHRL CONSIDERATIONS

By contextualizing PsyOps in the previous section, it can be understood how their conduct could raise human rights concerns. This is most clearly the case in relation to the possible impact on the right to freedom of opinion under IHRL, which provides for the freedom to hold opinions without interference.[30]

The UN's Special Rapporteur on the promotion of the rights of freedom of opinion and expression has asserted that *'coercive, involuntary or non-consensual manipulation of the thinking process to develop an opinion'* is a violation of the right to freedom of opinion. This includes *'techniques that allow State ... actors to access and influence the thoughts and opinions of people without their knowledge or consent'*.[31]

Violation of the right to freedom of opinion is not only an ethical issue; it could also have practical consequences. As the Special Rapporteur notes, in the context of armed conflict, disinformation about the location and nature of hostilities, the displacement of troops or population, or the existence and accessibility of safe areas and essential services could lead people *'to make wrong and dangerous decisions'*.[32]

Freedom of expression is also potentially engaged in the sense that it incorporates the right to seek and receive information.[33] The Special Rapporteur notes that the use of disinformation and other manipulation of online content could disrupt the free flow of information, which has been described as a *'critical element'* of the right.[34] This is particularly so in armed conflict, where the right to receive accurate, trustworthy information to inform decision-making becomes a *'survival right'*.[35]

In the NATO context, such concerns are to some extent mitigated by the policy constraints in NATO doctrine that PsyOps should be truthful and attributable. However, this will not necessarily be replicated in the doctrines of all state militaries.

A further concern is the right to privacy. This right provides that no one shall be subjected to arbitrary or unlawful interference with their privacy, family, home or correspondence, or to attacks upon their honour and reputation.[36] Privacy is generally

30 See, for example, International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 17, art 19 (ICCPR).
31 OHCHR, 'Disinformation and Freedom of Opinion and Expression: Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (13 April 2021) UN Doc A/HRC/4725, para 36 <https://www.ohchr.org/en/calls-for-input/report-disinformation> accessed 4 March 2024.
32 OHCHR (n 8) para 21.
33 ICCPR (n 30) art 19.
34 OHCHR (n 31) para 38.
35 OHCHR (no 8) para 5.
36 ICCPR (n 30) art 17.

understood to include information privacy as a derivative right.[37] The means available now to inform and employ PsyOps have greater implications for this right. The right may be engaged in the PsyOps context by intelligence collection on TAs, particularly as PsyOps become more targeted and noting the extent of information that NATO PsyOps doctrine states is required for TAA. It may also be engaged if private information is published about a civilian subject, particularly where it relates to sensitive aspects of someone's private life, such as their sexuality.

The difficulty in addressing human rights concerns as part of military training and planning for IAC is that the application of IHRL to armed conflict, and particularly IAC, is complex. As a matter of legal principle, human rights continue to apply in armed conflict, just as in peacetime.[38] However, a state is only bound by human rights obligations where that state has legal jurisdiction. The human rights regime was primarily designed for a peacetime context, to regulate the relationship between a state and individuals within that state, so its juridical scope was primarily territorial. While it is now accepted that in certain circumstances human rights can have extraterritorial application,[39] precisely when a state's extraterritorial jurisdiction arises is not entirely settled. The two most widely accepted bases for extraterritorial jurisdiction are where the agents of a state exercise physical power and control over individuals outside their territory and where a state has effective control over the territory of another state. A clear example of the first basis is military detention.[40] However, the law is far from settled in terms of what precisely constitutes effective control for the second basis, as every situation turns on its own facts. Even in full war fighting in IAC, where it could reasonably be assumed that any form of effective control is impossible, there is still a view that a state could have sufficient effective control to trigger IHRL obligations.[41]

If there is jurisdiction, the second step is to determine which human rights the state is bound to uphold. Do all rights apply? Again, there is no straightforward answer to this. For effective control cases, it would appear that which human rights are in scope is a matter of the level of control being exercised; that is, how 'effective' control actually is. At the highest end of this scale, where a state exercises a level of control akin to that in its national territory, the full range of human rights must be applied. However, in situations where control is more fragile, only those human rights that are

37 Robin Geiss and Henning Lahmann, 'Protection of Data in Armed Conflict' (2021) The Geneva Academy of International Humanitarian Law and Human Rights Working Paper, 8 <https://www.geneva-academy.ch/news/detail/ 452-three-papers-map-contentious-issues-related-to-the-application-of-international-law-to-military-cyber-operations> accessed 4 March 2024.

38 See, for example, *Legal Consequences of the Construction of a Wall* (Advisory Opinion) ICJ Rep 136, paras 106, 111.

39 See, for example, *Al-Skeini and Others v the United Kingdom* (GC) App no 55721/07 (ECtHR, 7 July 2011) para 149.

40 See, for example, *Hassan v the United Kingdom* (GC) App no 29750/09 (ECtHR 16 September 2014) para 76.

41 Marko Milanovic, 'Georgia v. Russia No. 2: The European Court's Resurrection of Bankovic in the Contexts of Chaos' (*EJIL: Talk!*, 25 January 2021) <https://www.ejiltalk.org/georgia-v-russia-no-2-the-european-courts-resurrection-of-bankovic-in-the-contexts-of-chaos/> accessed 4 March 2024.

'realistically relevant in the context' apply.[42] For the jurisdiction model based on state agent physical power and control over an individual, the state in question is under an obligation *'to secure to that individual the rights and freedoms … that are relevant to the situation of that individual'*.[43] So again, which rights are in scope will be context-dependent.

To date, the International Court of Justice and regional human rights courts have not considered the right to privacy or to freedom of opinion and expression in the context of armed conflict in any detail. It is therefore difficult to assess the level of control that would be needed for these rights to be in scope. Presumably it would be high, so as to have the resources and enforcement infrastructure in place to guarantee these rights. Even then, it is possible for the state's level of control to decrease again. In this case, the human rights applicable at any particular time could be in a state of constant flux.

If it is accepted that the rights to privacy and freedom of opinion and expression may be engaged in certain situations of IAC, what is the interaction between IHL and IHRL? The approach taken by the European Court of Human Rights (ECtHR) in *Georgia v Russia* (II)[44] was that the two regimes are applied concurrently, with each body of law mutually informing the interpretation of the other. One will only take precedence where there is a direct conflict between them that cannot be resolved other than by a policy decision to apply one over the other. A direct conflict between IHL and IHRL can certainly be envisaged. If all military PsyOps are required to be openly attributable and truthful or contain a caveat on their reliability, this may significantly impact their utility in the context of IAC. This direct conflict would then require a policy decision as to which body of law applies. This would almost certainly create uncertainty for military PsyOps personnel and lead to potential errors in the application of the legal framework.

It is not the intention of this paper to dismiss IHRL, but the reality is that when training military personnel, it is imperative to ensure that the legal parameters of their proposed activity are clear and to avoid context-based solutions. In contrast to the uncertainty about the scope and application of IHRL to IAC, IHL has no context-specific degrees of application. Accordingly, it may be preferable to consider whether the concerns about PsyOps raised by IHRL can be addressed and encompassed by IHL. This is the purpose of the next section.

42 Antal Berkes, *International Human Rights Law Beyond State Territorial Control* (CUP 2021) 42–43.
43 *Al-Skeini* (n 39) paras 138–140.
44 *Georgia v Russia (II)* (GC), App no 38263/08 (ECtHR, 21 January 2021).

4. THE IHL FRAMEWORK

Military PsyOps activity in IAC, just as any other military activity, must comply with IHL. The key foundational documents for IHL relevant to this paper are the four Geneva Conventions (GCI, GCII, GCIII and GCIV)[45] along with their first Additional Protocol (AP I).[46] At the time that these were written, their primary concern was to regulate the use of kinetic force, as this posed the greatest danger to civilians.

Against this background, it is unsurprising that IHL contains no specific provisions addressing the right to freedom of opinion and expression or the right to privacy. It is also unsurprising, given the generally non-kinetic nature of PsyOps, that reference to influence activity in IHL is limited. The key provision is Article 37(2) AP I, which deals with ruses of war, defined as acts that are intended to mislead an adversary or to induce them to act recklessly. It is widely accepted that PsyOps are ruses of war. The ICRC commentary to AP I specifically gives as an example of a ruse *'resorting to psychological warfare methods by inciting the enemy soldiers to rebel, to mutiny or desert'*.[47] State practice also supports the categorization of PsyOps as ruses of war; for example, the Australian Law of Armed Conflict Manual explicitly refers to PsyOps as *'legitimate ruses'*.[48] There is also academic support for this categorization[49]

Ruses of war are not prohibited, provided that they are not perfidious and do not infringe any other applicable legal rule. A perfidious act is one designed to make the adversary believe that they are entitled to receive or must grant protection under IHL, in order to exploit this confidence to capture, kill or wound the adversary.[50] The prohibition on perfidy provides only a limited means of regulating PsyOps from a civilian protection perspective, given its adversary focus and narrow scope, and will not be considered further here. There are, however, other rules of IHL that do regulate PsyOps from this perspective. Before moving on to look at these, one particular issue that frequently causes confusion must be considered: can PsyOps, as ruses of war, be directed at civilians?

45 Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 31 (GCI); Geneva Convention for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 85 (GCII); Geneva Convention Relative to the Treatment of Prisoners of War (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 135 (GCIII); Geneva Convention Relative to the Protection of Civilian Persons in Times of War (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 287 (GCIV).

46 Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol 1) (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3 (AP I).

47 Yves Sandoz and others (eds), *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (Martinus Nijhoff Publishers 1987) para 1521.

48 *Law of Armed Conflict* (Australian Defence Doctrine Publication 06.4, May 2006) para 7.1 <https://www.onlinelibrary.iihl.org/national-military-manuals/> accessed 4 March 2024.

49 Michael N Schmitt and Liis Vihul (eds), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (2nd edn, CUP 2017) 496.

50 AP I (n 46) art 37(1).

The military doctrine explored earlier appears to accept, in framing its potential TAs, that PsyOps may be directed at civilians. This is supported by the *Tallinn Manual*'s International Group of Experts (IGE), which agreed that *'psychological operations such as making propaganda broadcasts and dropping leaflets are permitted even if against civilians'*.[51] However, at first glance, this seems hard to reconcile with the IHL principle of distinction, which provides that civilians must not be the object of attack.[52]

The key point is that distinction only governs attacks, the main characteristic of which is the use of violence or force.[53] The US Law of War Manual addresses this issue directly. It states that the principle that military operations must not be directed against civilians does not prohibit military operations short of violence that are militarily necessary. It gives PsyOps as an example.[54] The *Tallinn Manual*'s IGE also agreed that non-violent operations, such as psychological cyber operations, do not qualify as attacks.[55]

However, while PsyOps may not directly involve the use of kinetic force, they may nevertheless cause violent effects. An example would be PsyOps that incite violent public disorder as a means of destabilizing the adversary state, resulting in civilian death and injury and damage to civilian property. The French Law of Military Operations Manual also addresses this issue. It refers specifically to the use of PsyOps to discredit someone. Where the consequences of such an operation could lead indirectly to the neutralization of the targeted individual by, for example, the adversary, the full kinetic targeting process must be applied.[56]

If violent effects are identified during operational planning as a reasonably foreseeable consequence of the PsyOp in question, the operation is in fact an attack. As an attack, it is subject to the principle of distinction and all other IHL provisions governing attack. This is the same approach applied to assess whether planned cyber operations reach the level of attack.[57]

It would seem unlikely that many military PsyOps will cause violent effects so as to render them attacks, given their non-kinetic nature. Nevertheless, this issue must still be considered in military planning. For military planning purposes, the question is whether PsyOps are reasonably expected to cause violent effects. Accordingly, even if not *prima facie* an attack, all proposed PsyOps should be subject to an assessment

51 Schmitt and Vihul (n 49) 421.
52 AP I (n 46) art 48.
53 ibid art 49(1); Sandoz and others (n 47) para 1875; Schmitt and Vihul (n 49) para 422.
54 US Department of Defense (n 12) para 5.2.2.1.
55 Schmitt and Vihul (n 49) 15.
56 Le ministère des Armées (n 13) para 8.1.3.1.
57 Schmitt and Vihul (n 49) 415, 416.

and authorization process, with legal support. This should involve a multi-stakeholder discussion on the proposed operation, working through its potential ramifications.

A further associated issue is whether data is an object for the purposes of targeting and therefore also subject to the principle of distinction that states that civilian objects, as well as civilians, must not be attacked. This is clearly relevant to PsyOps that seek to manipulate or delete data (for example, social media content). Academic opinion remains divided, but the majority of the *Tallinn Manual*'s IGE were of the view that data cannot be an object due to its intangible nature.[58] Using this approach, cyber-enabled PsyOps to manipulate or delete data, be it of a military or civilian nature, are not considered attacks in the absence of any accompanying physical damage or destruction.

Even if it is accepted that PsyOps activity will rarely be categorized as an attack, IHL does contain other more general provisions that protect civilians, directly or indirectly. These are considered below.

A. Obligation of Constant Care

Article 57(1) of AP I states that in the conduct of 'military operations', constant care shall be taken to spare the civilian population, civilians and civilian objects. The concept of military operations would appear to be broader than attack. This is based on the fact that the subsequent sub-article, Article 57(2), goes on to deal specifically with attacks. If the term 'military operations' is considered to be broader than attack, this would extend the reach of the obligation, enhancing the protection of civilians.

That military operations and attacks are distinct terms is supported by the ICRC commentary to AP I. This states that military operations should be understood to mean all the movements and activities carried out by armed forces related to hostilities.[59]

The UK Law of Armed Conflict Manual also supports a broad interpretation of military operations, pointing out that this term has *'a wider connotation than "attacks" and would include the movement or deployment of armed forces'*.[60] There is also academic support to extend the understanding of military operations to non-kinetic operations. It has been asserted that the term should encompass *'all informational operations necessary to support military activities including intelligence collection'*.[61] This would clearly capture PsyOps, including TAA.

58 Schmitt and Vihul (n 49) 437.

59 Sandoz and others (n 47) paras 1936, 2191.

60 *The Joint Service Manual of the Law of Armed Conflict* (UK Joint Service Publication 383, 2004) para 532 <https://www.gov.uk/government/groups/development-concepts-and-doctrine-centre#legal> accessed 4 March 2024.

61 Asaf Lubin, 'The Duty of Constant Care and Data Protection in War' in Laura A Dickinson and Edward Berg (eds), *Big Data and Armed Conflict: Legal Issues Above and Below the Armed Conflict Threshold*, Indiana Legal Studies Research Paper No. 473, 11 <https://ssrn.com/abstract=4012023> accessed 4 March 2024.

In terms of the precise scope of the obligation, the UK Law of Armed Conflict Manual explains the obligation as *'the commander will have to bear in mind the effect on the civilian population of what he is planning to do and take steps to reduce that effect as much as possible'*.[62] However, it remains unclear what harm civilians are meant to be spared from. Some doubt has been expressed as to whether the concept of harm in this context can be expanded beyond violent effects, noting the lack of supporting state practice.[63] However, there is an opposing academic view that the types of harm covered are not limited in this way and have a broader scope.[64]

Notwithstanding opposing views on the concept of harm, the purpose of this provision is clear: to protect civilians. The obligation of constant care should be considered as part of the general assessment and authorization process for PsyOps proposed earlier. The provision provides a valuable bridge between PsyOps that are attacks and those that are military operations, and it provides a hook to consider some of the broader questions about the protection of civilians raised by IHRL. A common-sense approach should be applied to work through the potential implications of a proposed operation for civilian protection and to consider mitigation and alternative courses of action.

B. Prohibition on Terrorizing the Civilian Population

Article 51(2) of AP I prohibits attacks or threats of attack that are primarily intended to spread terror, or extreme fear, among the civilian population. In the PsyOps context, threats of violence are likely to be most relevant. The French Law of Military Operations Manual gives as an example threatening civilians with attack if they do not leave an area as instructed.[65] Simply disseminating PsyOps products on social media with terrorizing content, for example, footage of an actual attack, with no articulated threat would not meet the threshold, although it may trigger the obligation to take constant care.[66]

This raises the question of whether a threat of attack may be implied. For example, leaflets dropped by the Israeli Defence Force on Gaza residents in October 2023 warned the civilian population to leave the area immediately or risk their lives, ostensibly constituting a precautionary measure under Article 57 (2) AP I. However, the leaflets added that anyone choosing not to evacuate may be considered an accomplice in a terrorist organization.[67] This could be construed as an implied threat of violence against civilians who choose not to leave their homes, to terrorize them into compliance.

62 *UK Manual of the Law of Armed Conflict* (n 60) para 5.32.1.
63 Geiss and Lahmann (n 9) 13.
64 Lubin (n 61) 14, 15.
65 Le ministère des Armées (n 13) para 8.1.3.1.
66 Geiss and Lahmann (n 9) 12.
67 Donatella Rovera, 'Amnesty International, Israel/OPT: Israeli Army Threats Ordering Residents of Northern Gaza to Leave May Amount to War Crimes' (*Amnesty International*, 25 October 2023) <https://www.amnesty.org/en/latest/news/2023/10/israel-opt-israeli-army-threats-ordering-residents-of-northern-gaza-to-leave-may-amount-to-war-crimes/> accessed 4 March 2024.

The prohibition on terrorizing civilians should be specifically considered during military planning for PsyOps, as part of the suggested general assessment and authorization process. Proposed PsyOps products should be scrutinized during military planning to ensure that they cannot be construed to contain either an implied or express threat of attack against civilians. This includes how PsyOps products are disseminated to ensure that products designated for an adversary TA, which may directly threaten violence, are not inadvertently disseminated to civilians.

Of note, PsyOps should not use images or video footage of identifiable dead bodies to instil fear among civilians. Several IHL provisions mandate that the dead must be respected and protected during armed conflict.[68]

C. Obligation of Humane Treatment

Common Article 3 (CA3) of all four Geneva Conventions sets a benchmark for the humane treatment of persons taking no active part in hostilities in armed conflict. Outrages upon personal dignity, in particular humiliating and degrading treatment, are specifically prohibited.

CA3 is supplemented by specific provisions in the Conventions. For civilians, this is Article 27 GCIV. Civilians who find themselves *'in the hands of'*[69] the adversary, either during hostilities or in occupation are entitled to humane treatment and to protection against insults and public curiosity. They are entitled to respect for their persons, their honour and their family rights. This mirrors the protections provided to prisoners of war under GCIII.

The ICRC commentary on Article 27 states that respect for someone's person encompasses both intellectual and physical aspects. Accordingly, individual persons' names or photographs, or aspects of their private lives must not be publicized. Individuals must also not be slandered or insulted, nor should any other action be taken that may affect their reputation.[70] This would appear to prohibit PsyOps that seek to discredit or undermine a civilian subject, for example, by releasing sensitive private information about them. Not only could this be humiliating, but it could also, depending on the information released, place them in physical danger.

There is some dispute about the scope of Article 27. It undoubtedly protects civilians in the physical control of an adversary, such as detainees, and civilians living in occupied territory. However, there appears to be a gap in terms of general coverage before military occupation has been established – that is, during the war-fighting phase.[71]

68 GCIV (n 45) art 16(2); AP I (n 46) art 34(1).

69 GCIV (n 45) art 4.

70 Jean S Pictet (ed), *Geneva Convention Relative to the Protection of Civilian Persons in Time of War. Geneva, 12 August 1949: Commentary* (ICRC 1958).

71 Kubo Mačák and Mikhail Orkin, 'Who Is Protected by the Fourth Geneva Convention? The Case of Civilians in Invaded Territory' (*Lieber Institute West Point*, 15 August 2022) <https://lieber.westpoint.edu/who-is-protected-civilians-invaded-territory/> accessed 4 March 2024.

Nevertheless, given that Article 27 is an expansion of CA3, it should be read in the spirit of that provision and considered during military planning for PsyOps, as part of the suggested general assessment and authorization process. This is particularly so as its provisions may go some way to addressing concerns about privacy.

D. Obligation to Respect and Protect Medical Services

Several IHL provisions mandate that medical personnel and units, both military and civilian, must be respected and protected and may not be attacked.[72] It has been suggested that this protection includes personal medical data, such as patient records, as well as any other data *'belonging to medical units and their personnel'*.[73] This approach would clearly prohibit the cyber-enabled extraction and use or manipulation of personal medical data for PsyOps purposes from a medical unit or a healthcare professional.

According to the *Tallinn Manual*'s IGE, 'respecting' medical services and infrastructure implies a state's obligation to refrain from carrying out operations that impede or prevent medical personnel from performing their medical functions or that otherwise adversely affect the humanitarian functions of medical personnel.[74] It is clear that the extraction and use/manipulation of personal medical data could be detrimental to the functioning and humanitarian purpose of a medical unit, in terms of both patient trust and medical treatment.

This also ties in with broader concerns about the use of disinformation to distort information vital to securing human needs, including medical services. This was seen in the COVID pandemic, with people declining to be vaccinated as a result of disinformation about the health risks of vaccines.[75] In an armed conflict scenario, disinformation could be used to discredit medical units and personnel in order to discourage people from using their services, with implications for their physical well-being. From a military PsyOps perspective, such operations would be strictly prohibited. Particular attention must be paid to whether planned PsyOps may affect medical services during the recommended assessment and authorization process.

5. CONCLUSION

Despite the observation that PsyOps has *'basically only the prohibition of perfidy as a constraint'*,[76] it can be seen that there are in fact a number of provisions in IHL

72 See, for example, GCI (n 45) arts 19, 24, 25, 35 and 36; AP I (n 46) arts 12, 15, 21–24 and 26.
73 Schmitt and Vihul (n 49) 515.
74 ibid 514.
75 Ronin Emmott, 'Russia, China Sow Disinformation to Undermine Trust in Western Vaccines: EU' (*Reuters*, 28 April 2021) <https://www.reuters.com/world/china/russia-china-sow-disinformation-undermine-trust-western-vaccines-eu-report-says-2021-04-28/> accessed 4 March 2024.
76 Geiss and Lahmann (n 9) 3.

that are relevant to and regulate PsyOps from a civilian protection perspective. While IHL does not specifically address the human rights concerns discussed earlier, these provisions provide the mechanism for these broader concerns to be considered as part of military planning.

Accordingly, in the context of training and planning for PsyOps in IAC, the military focus should remain on IHL as the overarching legal framework. All proposed PsyOps should be subject to an assessment and authorization process, in the course of which the wider implications of a particular operation for civilians and the civilian population should be considered. This process should be legally supported, in the same way as the targeting process. For NATO and its member states, this process is already envisaged in NATO doctrine and simply requires development, acknowledging the increased complexity of PsyOps in light of advances in ICT.

While some may be sceptical as to whether states will be willing to apply an expansive interpretation of IHL to accommodate wider civilian protection concerns, it should be noted that NATO doctrine, which demonstrates general consensus between 32 states, already addresses some of these concerns. It mandates that NATO PsyOps should be generally truthful and attributable, which is not in fact a requirement of IHL, and expressly requires a broad consideration of their potential effects.

Above all, a degree of reality and pragmatism needs to be applied when considering training and planning for military PsyOps in IAC. The legal regime must be articulated clearly so that military personnel know the parameters they are operating within; there is no room for lengthy academic debate. An overemphasis on IHRL, given its uneven and contextual application, may cause confusion. The focus should be on strengthening IHL, emphasizing its essential humanitarian purpose and using its existing provisions to ensure that the protection of civilians is a central consideration in training and planning for military PsyOps in IAC.

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

Reflections on the Afterlife: Which Rules Govern the Post-Occupation Retention and Use of Personal Data Collected by the Military?

Tatjana Grote
PhD Candidate
School of Law
University of Essex
Colchester, United Kingdom
t.grote@essex.ac.uk

Abstract: The collection of personal data by the military has become a commonplace practice in situations of military occupation. However, there has been barely any scholarly engagement with the post-occupation afterlife of such data. This paper explores the question of which legal regime governs the retention and use of personal data collected during a military occupation once the former occupying power is no longer in a position to exercise territorial control. It argues that as long as physical control over territory or a person is required to establish the extraterritorial applicability of international human rights law (IHRL), the pertinent IHRL treaty provisions will no longer be applicable to personal data collected during an occupation once a former occupying power ceases to exercise territorial control. While the law of occupation is traditionally equally premised on physical control, there is a convincing case for the continuous applicability of certain international humanitarian law (IHL) provisions based on a functional approach to the law of occupation. Although the relevant IHL norms do not protect privacy as an end in itself and hence offer only very limited protection, they might constitute the only available legal safeguard regarding the post-occupation retention and use of personal data if and to the extent that data protection law and IHRL are inapplicable.

Keywords: *personal data, international humanitarian law, international human rights law, military occupation, post-occupation law*

1. INTRODUCTION

When the United States Armed Forces left Afghanistan, they left behind, *inter alia*, handheld interagency identity detection equipment (HIIDE),[1] which is used to collect biometric data, identify individuals and access contextual information on them.[2] Part of this data has now most likely become accessible to the Taliban, putting many Afghans at a considerable risk of retribution.[3] This example is emblematic of two relatively recent developments: First, personal data collection has become an integral part of military control.[4] Second, the afterlife of such data can be highly relevant to the well-being of civilians.

So far, the academic debate on data protection and armed conflict has focused either on the rules relating to the conduct of hostilities[5] or the collection of data during a belligerent occupation.[6] Issues related to the post-conflict or post-occupation storage and use of personal data collected by the military, on the other hand, have received barely any attention.

This lack of scrutiny is concerning as personal data is a highly sought-after resource that equips a belligerent party with a temporally unlimited form of influence. Even long after the end of an occupation, the data collected by a belligerent party could become a threat to the data subject when falling into the hands of malign actors. Equally, the belligerent party itself might continue using the data it collected to exercise remote control over the behaviour of data subjects, such as by threatening to publish harmful information.

1 Leah West, 'Face Value: Precaution versus Privacy in Armed Conflict' in Asaf Lubin and Russell Buchan (eds), *The Rights to Privacy and Data Protection in Times of Armed Conflict* (NATO CCDCOE Publications 2022) 132–133.

2 ibid 133.

3 Human Rights Watch, 'New Evidence That Biometric Data Systems Imperil Afghans' (30 March 2022) <https://www.hrw.org/news/2022/03/30/new-evidence-biometric-data-systems-imperil-afghans> accessed 27 November 2023.

4 Asaf Lubin, 'The Rights to Privacy and Data Protection under International Humanitarian Law and Human Rights Law' in Robert Kolb, Gloria Gaggioli and Pavle Kilibarda (eds), *Research Handbook on Human Rights and Humanitarian Law* (Edward Elgar Publishing 2022) 465; Marten Zwanenburg, 'Know Thy Enemy: The Use of Biometrics in Military Operations and International Humanitarian Law' (2021) 97 International Law Studies 1405, 1411–1413.

5 See eg West (n 1); Tim McCormack, 'International Humanitarian Law and the Targeting of Data' (2018) 94 International Law Studies 222; Kubo Mačák, 'Military Objectives 2.0: The Case for Interpreting Computer Data as Objects under International Humanitarian Law' (2015) 48 Israel Law Review 55; Heather A Harrison Dinniss, 'The Nature of Objects: Targeting Networks and the Challenge of Defining Cyber Military Objectives' (2015) 48 Israel Law Review 39.

6 Rohan Talbot, 'Automating Occupation: International Humanitarian and Human Rights Law Implications of the Deployment of Facial Recognition Technologies in the Occupied Palestinian Territory' (2020) 102 International Review of the Red Cross 823; Omar Yousef Shehabi, 'Emerging Technologies, Digital Privacy, and Data Protection in Military Occupation' in Russell Buchan and Asaf Lubin (eds), *The Rights to Privacy and Data Protection in Times of Armed Conflic*t (NATO CCDCOE Publications 2022).

This paper provides a preliminary analysis of one of the most crucial legal questions raised by this phenomenon: which law governs the post-occupation retention and use of civilian personal data[7] collected by the military?

This particular focus reflects the gap in the literature highlighted above as well as the practicalities of data collection: Collecting personal data will often only be possible and conducive to the security needs of a belligerent party once the armed forces repeatedly encounter civilians outside of a situation of hostilities, such as when exercising somewhat permanent control.[8] Moreover, the following discussions will centre around the law governing international armed conflict (IAC). While data collection might also be increasingly relevant in non-international armed conflict (NIAC),[9] the risk of a significant gap in protection is far smaller since post-conflict state conduct would be governed by IHRL.

Note that, in principle, domestic data protection law could apply to the phenomenon this paper is concerned with. However, if such laws exist,[10] activities relevant to national security will frequently be excluded from the scope of peacetime data protection legislation.[11] Moreover, domestic data protection law might not apply to foreign armed forces. Hence, without prejudice to the particularities of a specific body of domestic law, IHL and IHRL will most likely be the two main legal frameworks of relevance to military data collection and use when connected to an international armed conflict.

This paper argues that there is a convincing case for the continuous applicability of certain IHL provisions. While the protection provided by IHL will most likely be very minimal, since IHRL will no longer be applicable once a belligerent party ceases to exercise physical control over a person or territory, it might constitute the only available legal safeguard regarding the long-term storage and use of personal data collected in situations of military occupation.

7 The term personal data is understood as in art 4(1) of Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119.

8 Zwanenburg (n 4) 1408.

9 Myanmar is one example of this tendency; International Crisis Group, 'Myanmar's Military Struggles to Control the Virtual Battlefield' (18 May 2021) <https://www.crisisgroup.org/asia/south-east-asia/myanmar/314-myanmars-military-struggles-control-virtual-battlefield> accessed 28 October 2023.

10 Note that Afghanistan, for instance, did not have a data protection law at the time; Human Rights Watch (n 3).

11 See GDPR art 2(2)(a), (b) and (d). Some have also questioned the applicability of data protection law during armed conflict in general; see Robin Geiss and Henning Lahmann, 'Protection of Data in Armed Conflict' (2021) 97 International Law Studies 556, 568.

2. DOES IHRL APPLY TO THE POST-OCCUPATION RETENTION AND USE OF MILITARY PERSONAL DATA?

Regarding issues arising in the aftermath of conflict, IHRL might be a more appropriate point of departure than IHL. In NIAC, state parties act within their own territory, where IHRL would certainly be applicable *ratione loci*. In IAC, on the other hand, IHRL can only be relevant if and to the extent that it binds states when they are acting extraterritorially.[12] Consequently, there is a need to assess whether IHRL applies to the long-term storage and use of data collected during military occupation.

A. How Does the Post-Occupation Retention of Personal Data Relate to the Accepted Models of the Extraterritorial Applicability of IHRL?

It is by now fairly accepted that IHRL will apply extraterritorially when a state exercises effective control over foreign territory or when a state agent 'exercises control and authority over an individual'.[13] Note that the effective control over territory standard is considered to require a lower level of control than Article 42 Hague Regulations (HR).[14] Hence, most would agree that as long as and where territory is occupied within the meaning of Article 42 HR, IHRL will be applicable *ratione loci*.[15] What happens, however, once a belligerent occupation ends?

The underlying issue in this respect is the following: data-facilitated control can impact an individual's enjoyment of human rights outside of the acting state's territory without that state exercising any kind of physical control over that individual. So far, both above-mentioned tests have been applied exclusively in situations where the control or authority in question was physical. While the effective control-test quite clearly refers to territory, the state agent authority-test could be interpreted to encompass purely virtual control over an individual. However, there is no jurisprudence confirming such an interpretation in the context of digital privacy. Consequently, it seems fair to assume that the established theories of extraterritorial applicability of IHRL currently do not encompass situations in which control is purely virtual.

12 Note that this would not necessarily be the case if the relevant rights – in this case, the right to (digital) privacy – were part of customary international law. However, given the lack of uniformity when it comes to the precise content and protection of a right to digital privacy, this author agrees with others who do not consider this to be the case. See Eliza Watt, *State Sponsored Cyber Surveillance: The Right to Privacy of Communications and International Law* (Edward Elgar Publishing 2021) 141.

13 *Al-Skeini and Others v United Kingdom* [2011] ECtHR [GC] 55721/07 [74]. See also UN Human Rights Committee, 'General Comment No. 31' (29 March 2004) UN Doc CCPR_C_21_Rev.1_Add.13-EN para 10.

14 Hanne Cuyckens, *Revisiting the Law of Occupation* (Brill Nijhoff 2017) 174; Tom Ruys and Sten Verhoeven, 'DRC v. Uganda: The Applicability of International Humanitarian Law and Human Rights Law in Occupied Territories' in Roberta Arnold and Noëlle Quénivet (eds), *International Humanitarian Law and Human Rights Law: Towards a New Merger in International Law* (Brill Nijhoff 2008) 179.

15 Noam Lubell, 'Human Rights Obligations in Military Occupation' (2012) 94 International Review of the Red Cross 317, 319.

B. Alternative Approaches to Establishing the Applicability of IHRL to Personal Data

The United Nations Office of the High Commissioner for Human Rights (OHCHR) has suggested that the applicability of IHRL can be determined by answering the question of which state physically controls the location of the server or device on which the data is stored.[16] This approach is unconvincing, however, since states could simply eschew their IHRL obligations by storing data on a server or device they do not have physical control over.[17] Moreover, from a practical perspective, it might be nearly impossible to identify the physical location of the data in question, especially if such data has been copied to multiple servers or devices.

Others have suggested moving away from control-based models altogether and endorsed, instead, a cause-effect approach to the extraterritorial applicability of IHRL. Whereby, whenever state behaviour has a negative effect on the enjoyment of human rights abroad, the negative human rights obligations of the respective state regarding the affected right or rights would be applicable.[18] The idea underlying this approach has gained prominent scholarly support[19] and has surfaced in General Comment No. 36.[20] While the details of such an interpretation would need to be worked out, this author considers it a suitable response to technological and environmental developments that can enable a substantial impact on the enjoyment of certain rights abroad without physical control over the affected individuals. However, the approach has not (yet) been endorsed by any international court in the context of the right to privacy.

In sum, physical control over territory or over a person remains a necessary requirement for the extraterritorial applicability of most human rights treaties. Consequently, under current international law, the long-term storage and use of data collected during a military occupation will, if at all, be governed by IHL.

16 UN Office of the High Commissioner for Human Rights, 'The Right to Privacy in the Digital Age: Report of the Office of the United Nations High Commissioner for Human Rights' (30 June 2014) UN Doc A/HRC/27/37 para 34.

17 See also Marko Milanovic, 'Surveillance and Cyber Operations' in Mark Gibney (ed), *The Routledge Handbook on Extraterritorial Human Rights Obligations* (Routledge 2021) 374–375.

18 ibid 373–375; Marko Milanovic and Michael N Schmitt, 'Cyber Attacks and Cyber (Mis)Information Operations during a Pandemic COVID-19' (2020) 11 Journal of National Security Law and Policy 247, 263; Marko Milanovic, *Extraterritorial Application of Human Rights Treaties* (Oxford University Press 2011) 209–219.

19 Yuval Shany, 'Taking Universality Seriously: A Functional Approach to Extraterritoriality in International Human Rights Law Borders and Human Rights' (2013) 7 Law & Ethics of Human Rights 47, 67–71. While Shany agrees with the general idea of moving away from control-based standards, he proposes a different, functional approach that emphasizes the intensity of the power and special legal relations between a state and an individual as factors to be taken into account when determining the existence and extent of extraterritorial human rights obligations.

20 UN Human Rights Committee, 'General Comment No. 36' (3 September 2019) UN Doc CCPR/C/GC/36 para 63.

3. DOES IHL APPLY TO THE POST-OCCUPATION RETENTION OF MILITARY PERSONAL DATA?

Generally, IHL applies once an armed conflict erupts or an occupation is established. It should be noted that certain rules come with specific applicability regimes. Hence, there is a need to evaluate which parts of IHL would apply to the post-occupation storage and use of military data.

A. Which Parts of IHL Are of Relevance, and Which Provisions Govern Their Applicability?

As explained above, military data has, so far, mostly been discussed from a conduct of hostilities perspective. Mass collection of civilian personal data, however, is most common in situations where the military exercises stable control.[21] The rules applicable to such situations would, thus, seem a natural starting point from which to address the question at hand.

Traditionally, the applicability of the law of occupation has been considered to be defined by Article 42 HR: 'Territory is considered occupied when it is actually placed under the authority of the hostile army'.

There has been some debate amongst scholars on whether the adoption of Geneva Convention IV (GCIV), by virtue of its Article 4(1), broadened the scope of the law of occupation to all situations in which a protected person finds themselves 'in the hands of' a belligerent party. However, the most recent International Committee of the Red Cross (ICRC) commentary on Geneva Convention III (GCIII) clarifies that it is Article 42 HR that defines the concept of occupation and that 'subsequent treaties, including the Geneva Conventions, have not altered this definition'.[22] Yet, the 'in the hands' standard still defines the personal scope of the general protections of GCIV. Both standards, therefore, are of relevance to this paper.

This paper argues that both Article 4(1) GCIV and Article 42 HR traditionally refer to the same type of control, namely physical control over territory or persons. In the case of Article 42 HR, this is evident from the wording of the provision, which links the concept of occupation with territory. Many would agree that the effective control required by Article 42 HR can only be established through the actual or

21 The most prominent examples of military data collection during armed conflict are the operations carried out by the United States in Afghanistan, Iraq and Israel in the occupied Palestinian territories.

22 ICRC, *Commentary on the Third Geneva Convention: Convention (III) Relative to the Treatment of Prisoners of War* (Cambridge University Press 2021) para 327.

potential presence of 'boots on the ground'.[23] Even those arguing that remote control can be sufficient to establish a belligerent occupation consider that it is the potential of *physical* control, as opposed to merely virtual control, that justifies such an interpretation.[24] This can be understood as a corollary of the idea that the authority required by Article 42 HR needs to be exclusive.[25] Even nowadays, it is difficult to imagine a situation where purely virtual influence would be sufficient to establish exclusive control.

The commentary to GCIV suggests that 'in the hands of' is similarly defined with reference to physical control. While the concept might be broader than that of occupation, the ICRC commentary nevertheless defines it with reference to control over territory.[26] When the commentary states that the concept 'need not necessarily be understood in the physical sense',[27] this clarifies that a person does not need to be under the direct, physical control of a belligerent party. However, at least traditionally, an individual would still need to be located in territory that is or can be brought under the physical control of a belligerent party to qualify as a protected person. Hence, it seems that it is rather the stability and potentially the level of control that distinguishes a situation in which a person is in the hands of a belligerent party from a situation of occupation – not the type of control.

B. How Does Data-Facilitated Control Challenge the Current Legal Framework?

As Lieblich and Benvenisti note, 'While the law of occupation assumed, traditionally, physical control on the ground, control can nowadays be exercised through various measures.'[28] Retaining sensitive personal data will allow a belligerent party to take precisely such measures of control that are not premised on ongoing physical dominance.[29] For instance, a belligerent party could threaten to publish or share certain sensitive data to coerce a data subject into a specific course of action. It could further use the collected data to pretend to be the data subject to the personal and/or economic detriment of the latter.

23 Orna Ben-Naftali, 'Belligerent Occupation: A Plea for the Establishment of an International Supervisory Mechanism' in Antonio Cassese (ed), *Realizing Utopia: The Future of International Law* (Oxford University Press 2012) 541; Yoram Dinstein, *The International Law of Belligerent Occupation* (Cambridge University Press 2009) 50. The presence of foreign troops has been considered a necessary component for the establishment of a belligerent occupation by the ECtHR in *Sargsyan v Azerbaijan* and *Chiragov v Armenia*; *Sargsyan v Azerbaijan* [2015] ECtHR [GC] 40167/06 [94]; *Chiragov and Others v Armenia* [2015] ECtHR [GC] 13216/05 [96].

24 Dieter Fleck, 'Occupation' in Dieter Fleck (ed), *The Handbook of International Humanitarian Law* (Oxford University Press 2021) 298; Cuyckens (n 14) 37; Tristan Ferraro, 'Determining the Beginning and End of an Occupation under International Humanitarian Law' (2012) 94 International Review of the Red Cross 133, 145.

25 Fleck (n 24) 197.

26 The commentary states that '"in the hands of" … simply means that the person is in territory which is under the control of the Power in question'; Jean Pictet and others, *Commentary on the Geneva Conventions of 12 August 1949, Vol IV* (ICRC 1958) 47. See also *Prosecutor v Vinko Martinović and Mladen Naletilić* (Judgment) IT-98-34 (31 March 2003) [208].

27 Pictet and others (n 26) 47.

28 Eliav Lieblich and Eyal Benvenisti, *Occupation in International Law* (Oxford University Press 2023) 221.

29 See also Lubin (n 4) 465.

As shown above, the applicability of IHL norms regulating the exercise of control over a person is delimited mainly in territorial terms. While the collection of data will frequently be effectuated through physical installations (e.g., CCTV) or during physical encounters (e.g., using HIIDE), personal information can be stored and used in any place once transformed into a computer-readable format.[30] This gives rise to a precarious situation: having collected sensitive data while being in physical control of a person or territory, a belligerent party might continue to hold some authority over individual data subjects, even without exercising physical control over them or the territory they are located on.

Thus, data-facilitated control does not fit squarely within the established mechanisms of determining the applicability of those IHL provisions dealing with the exercise of control over an individual as it de-territorializes and compartmentalizes control. The first is due to the virtual nature of data. The second is a consequence of the fact that sensitive data might allow the data holder to exercise significant control over certain aspects of the data subject's life, which nevertheless falls short of complete control over almost all aspects of life a state has when occupying territory.

C. How to Respond to These Challenges?

Does this mean that in the absence of effective control over territory, the law of occupation and the general protections of GCIV would not apply? Not necessarily. At least two avenues leading to the conclusion that IHL will continue to apply to data collected during situations of military control can be identified. While the first would require a novel and thus far unsupported interpretation of Article 4(1) GCIV, the latter has gained prominent scholarly support and could reasonably be extended to the situation discussed in this paper.

D. Interpreting Retaining Personal Data as Remaining in the Virtual Hands of a Belligerent Party

It has been averred that 'taken seriously, the "in the hands" test, as applied to individuals, could perhaps be extended to situations beyond actual physical contact with troops'.[31] Arguing in this vein, one might contend that the prominence of digital means in everyday life has increased to an extent where the term 'person' must be understood as capturing both the physical body of an individual as well as any digital representation thereof. The latter could remain in the hands of a belligerent party even when the physical person is not anymore. This contention relies on the idea that personal data can be conceptualized as a virtual extension of the physical civilian, with the two being linked by the concept of human dignity, which encompasses both physical and non-physical components of the human experience. However, some

30 For instance, portions of the data collected by the US in Afghanistan were reportedly stored on servers in West Virginia; Public Intelligence, 'Identity Dominance: The U.S. Military's Biometric War in Afghanistan' (21 April 2014) <https://publicintelligence.net/identity-dominance/> accessed 14 October 2023.

31 Lieblich and Benvenisti (n 28) 67.

conceptual clarification would be necessary: Would all personal data, such as emails exchanged with a gardening company on the topic of hiring a lawnmower, or only certain especially sensitive categories of data bring a person into the virtual hands of a belligerent party? Moreover, it seems unlikely that courts or states would consider such an approach to constitute *lex lata*.

E. Towards an Extension of the Functional Approach to Personal Data

Some have argued that certain obligations of occupying powers might remain applicable even when the (former) occupying power ceases to exercise effective control.[32] More specifically, if a former occupying power 'retain[s] key elements of authority or other important governmental functions, the law of occupation may continue to apply within the territorial and functional limits of such competences'.[33] While effective control over territory might, therefore, be a necessary condition to *trigger* the applicability of the law of occupation, it might not be required to sustain the applicability of some of its provisions.[34] Some consider this a dangerous fragmentation of the law of occupation, which would create legal uncertainty and 'entrust the occupying power with the ability to determine the extent of its own obligations'.[35] Others argue that without such an approach, a local population might be left 'bereft of any major legal protection' in situations where authority is shared between belligerent states, which stands in stark contradiction with the *telos* of the law of occupation.[36] The post-occupation retention of personal data constitutes one manifestation of such a scenario.

It should be noted that the asymmetry between the test for determining the beginning of the temporal scope of the law of occupation and the test for its end remains undertheorized.[37] Concerning the long-term storage of personal data, this is crucial: If we accept that the law of occupation remains applicable with respect to such data, would we not also have to argue that collecting personal data through purely digital means can trigger the law of occupation in the first place? This would come close to

32 Ferraro (n 24) 157–158; Dinstein (n 23) 301–302; Iain Scobbie, 'An Intimate Disengagement: Israel's Withdrawal from Gaza, the Law of Occupation and of Self-Determination' (2004) 11 Yearbook of Islamic and Middle Eastern Law 3, 30; Aeyal Gross, The Writing on the Wall: Rethinking the International Law of Occupation (Cambridge University Press 2017) 133–134. Note that Gross specifically proposes a functional approach. The other authors advocated for asymmetrical approaches more broadly.

33 International Committee of the Red Cross, 'International Humanitarian Law and the Challenges of Contemporary Armed Conflicts' (International Committee of the Red Cross 2015) Report 32IC/15/11 12 <https://www.icrc.org/en/document/international-humanitarian-law-and-challenges-contemporary-armed-conflicts> accessed 22 May 2019.

34 See eg Dinstein (n 23) 301.

35 Cuyckens (n 14) 41–42; Valentina Azarov, 'Disingenuous "Disengagement": Israel's Occupation of the Gaza Strip and the Protective Function of the Law of Belligerent Occupation' (*Opinio Juris*, 24 April 2012) <https://opiniojuris.org/2012/04/24/disingenuous-disengagement-israels-occupation-of-the-gaza-strip-and-the-protective-function-of-the-law-of-belligerent-occupation/> accessed 27 November 2023; Yuval Shany, 'Faraway, So Close: The Legal Status of Gaza after Israel's Disengagement' (2005) 8 Yearbook of International Humanitarian Law 369, 378.

36 Ferraro (n 24) 302.

37 Lieblich and Benvenisti (n 28) 75.

accepting that there can be something like a 'cyber occupation' – an idea that has been prominently and rightly rejected.[38]

However, there are good reasons for this asymmetry. First, IHL is already fragmented regarding the end of its applicability. Indeed, prisoners of war are protected by GCIII 'until their final release and repatriation'.[39] Additional Protocol I (AP I) explicitly states that its applicability extends until the release, repatriation or re-establishment of a person, including after the termination of an occupation.[40] Certain provisions of Additional Protocol II (AP II) apply to individuals deprived of their liberty 'for reasons related to [a] conflict' even if this deprivation begins after the end of the armed conflict.[41] What this implies is that once an armed conflict erupts, IHL applicability mirrors the existence of protective needs. As long as protected individuals are in need of protection as a result of an armed conflict, whether that conflict is ongoing or not, the relevant norms of IHL apply.

Second, a functional approach aligns well with the spirit of the law of occupation. Many would agree that the obligations of an occupying power are the flipside of the authority over foreign people or territory that it exercises during an occupation. The need to have a body of law like the law of occupation emanates from the fact that, as a result of an armed conflict, a belligerent party has gained control over a foreign population, which allows it to impact this population's life in ways that are different from the effects of military attacks. By virtue of the stable control over territory it gained through militarily ousting the previous sovereign, it has created a relationship of dependence between itself and the local population that is not just momentaneous. This fact of the dependence of the local population on the occupying power,[42] without regard for its legitimacy, lies at the heart of the law of occupation. Since it is the occupying power that, for whatever reason, actively decided to create such a dependence, it also bears the responsibility for the effects of actions made possible by this dependence for as long as it decides to maintain it. If it was the fact of having physical control over a person or territory that enabled the creation of a data-facilitated relation of (partial) dependence between the former occupying power and the data subject, it would seem in line with the spirit of the law of occupation that the relevant IHL provisions remain applicable.

Note that if this applies to the provisions of the law of occupation proper, by way of a *maiore ad minus* argument, the same reasoning will apply to the 'in the hands' standard, which arguably requires a lesser level of control. The idea would be similar

38 NATO CCDCOE (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (2nd edn, Cambridge University Press 2017) 543.
39 art 4 GCIII.
40 art 3(b) AP I.
41 art 2(3) AP II.
42 Lieblich and Benvenisti (n 28) 78; Yaël Ronen, 'Post-Occupation Law' in Carsten Stahn, Jennifer S Easterday and Jens Iverson (eds), *Jus Post Bellum: Mapping the Normative Foundations* (Oxford University Press 2014) 430–431.

to the one expressed above: as long as there are persons in need of protection as defined by Article 4(1) GCIV, the norms providing protection to such persons in their specific situation remain applicable. Note that this does not create the same outcome as adopting a 'virtual in the hands of' approach. While the former could be interpreted as extending the scope of GCIV to persons who have never been under the physical control of one of the belligerent parties, the functional approach would require the existence of physical control over territory at some point for the initial applicability of the provisions in question to be triggered.

What would this mean in practice? To the extent that there exist relevant IHL provisions, the pertinent rules would govern: (1) the passive storage and (2) the active use or sharing of personal data collected during a belligerent occupation.

4. WHICH NORMS OF IHL WOULD BE RELEVANT?

The question that naturally flows from the previous section is the following: which norms would remain applicable? As this paper is exclusively concerned with determining the applicable law to the post-occupation afterlife of personal data collected by the armed forces, not the application of such norms, this analysis will focus on which rules would be *prima facie* relevant and hence applicable under the functional approach set out above.[43]

A. Rules Related to the Maintenance of Public Order and Security – Article 43 HR

To the extent that data-enabled control allows a foreign state to influence the public order, civil life or security in the former occupied territories, Article 43 HR would be pertinent.

There are two issues to address here. Can the protection of abstract values such as privacy ever be necessary to maintain public order, civil life and security? This will most likely depend on the facts of a specific case. If a data leak or data-facilitated action taken by the occupying power itself was a *sine qua non* condition for social disarray or physical harm to protected persons at the hands of non-state actors, this might, indeed, be considered precisely the kind of outcome that a former occupying

43 Note that this section will not discuss property-related provisions. As suggested in the discussions of the Tallinn Manual experts, the idea that data can be considered property remains a minority position; NATO CCDCOE (n 38) 550. However, should personal data be understood as property by the international community in the future, those provisions dealing with property during belligerent occupation would become further relevant. For a discussion of how to apply the existing property protection rules of IHL to data, *see* Eric Talbot Jensen and Laurie R Blank, 'LOAC and the Protection and Use of Digital Property in Armed Conflict' in Russell Buchan and Asaf Lubin (eds), *The Rights to Privacy and Data Protection in Times of Armed Conflict* (NATO CCDCOE Publications 2022).

power must protect data subjects from, to the extent possible.[44] However, it would be necessary to show that data-enabled conduct actually impacted civil life, public order or security. In the absence of such effects, Article 43 HR will not be pertinent.

A note of caution is in order: considering Article 43 HR pertinent could be interpreted as implying that all Article 43 HR-based positive duties apply – that is, there is a general duty to maintain civil life, public order and security. This would certainly be absurd, as holding personal data gives the former occupying power only very limited control over public order and civil life. Consequently, obligations based on Article 43 HR would be limited to those areas of public order and civil life that can actually be controlled by the occupying power. Unless the occupying power secures very specific data, such as sensitive data on all public servants, it will most likely not be able to control entire areas of public order or civil life. Moreover, it is unlikely that positive obligations would be applicable. However, a case-by-case assessment would be necessary to determine the precise scope of the obligations arising from Article 43 HR.

B. Rules Related to Fundamental Rights of Protected Persons – Article 27 GCIV

An article that would undoubtedly be relevant is Article 27 GCIV. Some have even proposed that the right to privacy could be read into the duty to respect the person, honour and family rights of protected persons.[45] However, the doctrinal permissibility and desirability of such an interpretation are far from obvious. As confirmed by the ICRC commentary on Article 27(1) GCIV, the obligations set out by the provision are absolute and cannot be overridden by security concerns.[46] If one interprets Article 27(1) GCIV to implicitly encompass the right to privacy, then this would either mean that guarantees protected by Article 27(1) GCIV can be limited or that any measure affecting a person's private life was unlawful. The first is a slippery slope, while the second sits uneasily with many provisions allowing for interferences with other rights, such as the right to liberty.[47] To argue that the right to privacy in general is protected in a more absolute manner seems arbitrary. For these reasons, this author is of the view that Article 27(1) GCIV cannot be read as establishing a general right to privacy under IHL.

However, specific uses of data might violate the duty to respect the person and honour of individuals in the hands of a state party. The concept of the *person* has been

44 Note that the International Court of Justice's reference to 'all the measures in [the OP's] power' leaves no doubt that the obligation is one of conduct, not of result; *DRC v Uganda* [2005] International Law Reports 168 (International Court of Justice) 178–179.

45 Talbot (n 6) 835. See also Shehabi (n 6) 98.

46 Pictet and others (n 26) 207; Yutuka Arai-Takahashi, *The Law of Occupation: Continuity and Change of International Humanitarian Law, and Its Interaction with International Human Rights Law* (Brill Nijhoff 2009) 275; Zwanenburg (n 4) 1422.

47 See art 78 GCIV. The ICRC commentary even states that the right to liberty was not included in art 27(1) GCIV precisely because it can be restricted; Pictet and others (n 26) 201.

considered as a broad one that, among other things, encompasses the 'intellectual integrity of human persons'.[48] The ICRC commentary on Article 27(1) GCIV specifies that respect for the intellectual integrity of a person requires that 'individual persons' names or photographs, or aspects of their private lives must not be given publicity'.[49] In light of this, Article 27(1) GCIV can be considered pertinent when data related to the private life of a person, photographs, or names are published, even if this is done after the end of the armed conflict or occupation.

Furthermore, Article 27(1) GCIV is relevant when personal data collected by the former occupying power would put protected persons at risk of violence, insults or threats of either, as well as when it exposes them to public curiosity. A crucial question concerns the precise meaning of the latter term. Does it refer to publishing information in fully public outlets only, or would it also include semi-public spaces (e.g., a shared database accessible to hundreds or thousands of third-party agents)? The ICRC commentary on GCIII states that '"public" should be interpreted as referring to anyone who is not directly involved in handling … prisoners of war, including other members of the Detaining Power'.[50] Whether a similar reasoning applies to Article 27(1) GCIV is one of the questions that will need to be addressed by the forthcoming updated commentary to GCIV, but this author sees no reason why it would not.

C. Rules Related to Specifically Prohibited Practices – Articles 31 and 33 GCIV

The mere threat of publishing sensitive information against the will of the data subject to obtain information from it could amount to moral coercion within the meaning of Article 31 GCIV. As specified in the ICRC commentary, 'the prohibition laid down in this article is general in character and applies to both physical and moral forms of coercion'.[51] Moreover, the pressure exerted can be 'direct or indirect, obvious or hidden'.[52] It is not difficult to imagine a situation where the threat of publishing or sharing sensitive data against the will of the data subject could be used to coerce an individual into sharing certain information. In such a scenario, Article 31 GCIV could provide legal protection.

If the former occupying power were to exercise data-facilitated control (e.g., by publishing or sharing sensitive data) over an entire group, this might further be considered a measure of intimidation or collective punishment pursuant to Article 33(1) GCIV. Note that 'penalties' has been interpreted as not only referring to penal sanctions but to 'penalties of any kind'.[53]

48 Arai-Takahashi (n 46) 271 (fn 37).
49 Pictet and others (n 26) 201.
50 ICRC (n 22) para 1624.
51 Pictet and others (n 26) 219.
52 ibid.
53 ibid 225.

In sum, unless one adopts a very broad interpretation of Article 27(1) GCIV, IHL does not protect privacy as an end in itself. The protection provided by IHL is, therefore, very limited and only shields data subjects from outcomes significantly affecting their dignity or physical or mental well-being, as well as from intimidation, collective punishment or coercion. Therefore, IHL is not necessarily well-suited to protecting the privacy of individuals after the end of a military occupation. Yet, if and to the extent neither IHRL nor domestic data protection laws are applicable, the mostly negative obligations that can be derived from IHL might constitute the only legal safeguards available.

5. CONCLUSION

This paper has explored the question of which legal regimes govern the long-term storage and use of data collected during military occupation. It argued that certain norms of IHL will remain applicable after the end of the occupation. While this author considers this view the most convincing among alternatives, others might reasonably come to different conclusions. This paper argued that as long as IHRL does not apply extraterritorially in situations where a state party has physical control over neither territory nor a person, IHL might provide the only legal safeguard available. If the view that IHRL applies extraterritorially whenever state conduct negatively impacts the enjoyment of rights abroad gains more support, the situation would need to be reassessed. However, so long as this is not the case, IHL will most likely be the only source of international legal restrictions on the retention and use of data collected during a belligerent occupation.

This paper did not seek to explore the applicable substantive norms in full detail. Rather, it is intended to be a conversation starter. While personal data protection might seem like the last thing to worry about in a situation of armed conflict or occupation, and also after it, this sentiment is a dangerous one. Armed conflict increasingly involves the targeting of individual persons as well as the social fabric of societies, including through individual humiliation, discreditation and sowing societal discord. Personal data constitutes one of the most valuable resources in this type of warfare. International lawyers, military and civil-society actors should, therefore, engage in a constructive dialogue on how to protect civilians not only from death, injury and destruction but also from humiliation, intimidation, coercion and other data-facilitated harms, both during and after armed conflict and military occupation.

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)
2024 © NATO CCDCOE Publications, Tallinn

Unity or Coherence: Shaping Future Civil-Military Intelligence Collaboration in the Cyber Domain

Neil Ashdown
Centre for Doctoral Training in Cyber Security
Royal Holloway, University of London
London, United Kingdom
neil.ashdown.2019@live.rhul.ac.uk

Abstract: Western militaries and governments are coming to recognize that they can only address the challenges of cyber defence by working with private sector cyber intelligence actors. It is almost certain that the involvement of the private sector in cyber defence will continue to expand in Western societies. Governments and militaries will therefore face the challenge of adapting to a world where – in the cyber domain – the state is not the sole provider of intelligence or security. Looking over the horizon, the paper considers two scenarios for the future of intelligence collaboration in the cyber domain within Western societies. In both scenarios, states aim to promote collaboration between industry, government, and the military, but the methods used differ. In the first scenario, Western states impose a unified structure for collaboration. In the second scenario, states aim for coherence, with governments seeking not to impose unity but to manage difference. The paper argues that the second approach is likely to be more successful but that it will require both a willingness to accept change and people able to act as translators between different organizations.

Keywords: *intelligence, civil, military, unity, coherence, collaboration*

1. INTRODUCTION

Western militaries and governments are coming to recognize that they can only address the challenges of cyber defence by working with private sector cyber intelligence actors. Deputy Commander UK Strategic Command Lt. Gen. Tom Copinger-Symes said in a September 2023 interview that in the cyber domain the UK military works with industry 'literally the whole time'. Lt. Gen. Copinger-Symes also emphasized the value of private-sector cyber intelligence for the military:

> We're very proud of how much cyberthreat data we gather as Defence, but that's tiny compared with what Microsoft gathers every day of the week … I mean, it's awesome, the scale they work at. (Martin 2023)

His comments underline two interconnected dynamics: the emergence of parts of the private sector as cyber intelligence actors and the importance of private sector cyber intelligence for modern militaries. Together, these dynamics raise the increasingly urgent question of how militaries can most effectively collaborate with industry.

The focus of this account is on the UK and the US, but a key argument of this paper is that cyber intelligence is a highly transnational practice, with the salient lines of division existing not at the country level but between groups of countries that share core values. This position has been made explicit by technology companies with advanced cyber intelligence capabilities responding to the Russia–Ukraine war (Smith 2022; Landau 2022). This account is therefore offered as a high-level description of the functioning of commercial cyber intelligence as it will be encountered by NATO militaries.

In the twenty-first century, parts of the private sector, through their access to data and their technical capabilities, have become important intelligence actors in the cyber domain (Work 2020; 2023; Zegart 2020; Healey and Korn 2019; Warner 2014). It is beyond the scope of this paper to debate whether this marks an unprecedented shift – in which, as Matthew Hurley argues, private companies can now 'conduct activities formerly the exclusive province of a state's security apparatus' – or, instead, the continuation of a long tradition of intelligence activity beyond the state (Hurley 2012, 18–19; Andrew 2018). What is key for this paper is that some of the activities that private sector organizations describe as 'cyber intelligence' are genuine intelligence activities rather than activities that approximate intelligence activity understood to be the preserve of state actors (Lindsay 2020; Stout and Warner 2018). The significance of this development remains underappreciated in discussions of cyber policy and strategy (Work 2023). It has also provoked important conceptual questions for practitioners in the private sector (Guerrero-Saade 2015).

This paper first describes private sector cyber intelligence and the networks of public and private actors that produce, exchange, and consume cyber intelligence. It then looks over the horizon, setting out two scenarios for the future organization of cyber intelligence collaboration between militaries and the private sector. The key difference will be between approaches that emphasize unity and those that emphasize coherence. These scenarios are followed by a discussion and a conclusion.

2. COMMERCIAL CYBER INTELLIGENCE

Organizational definitions of 'cyber intelligence' vary widely across the public and private sector (Work 2020; Bonfanti 2018; Bellaby 2016). These differences are a predictable consequence of organizations seeking to adopt definitions that align with their own histories, working practices, and ambitions (Slayton 2021; Lindsay 2020; Wiener 2016). These tendencies can be seen in various NATO member militaries' past attempts to integrate 'cyber intelligence' as a term into the established terminology of the intelligence collection disciplines, analytic practices, and military functions (Wiener 2016; Mattern et al. 2014; Hurley 2012; Neal-Hopes 2011). These challenges are not unique to the public sector; private sector organizations face similar challenges in defining cyber intelligence, while commercial pressures and marketing requirements shape how the term is applied (Work 2020; Bonfanti 2018).

Bonfanti offers a deliberately broad definition of cyber intelligence that is organizationally agnostic:

> [The term] cyber intelligence is used to convey the idea of widely scoped and better qualified knowledge of actual or potential events regarding cyberspace that may endanger an organization. (Bonfanti 2018, 107)

A broad definition can encompass the heterogeneity of cyber intelligence as a phenomenon (Work 2023; 2020; Kalkman and Wieskamp 2019). This heterogeneity is unsurprising given that, as JD Work argues, commercial cyber intelligence emerged in a 'complex and varied environment' (Work 2020, 279). It can be observed at the level of the people who work in private sector cyber intelligence, the range of cyber intelligence companies and functions, and the sharing networks that those people and organizations form, as well as in the role of cyber intelligence within the wider cyber ecosystem. This heterogeneity is important – it is a feature of commercial cyber intelligence rather than a bug.

Heterogeneity can be seen in the range of backgrounds and educations of people who work in commercial cyber intelligence. Former government and military personnel are

a crucial recruiting pool for commercial cyber intelligence (Healey and Korn 2019; Petersen and Tjalve 2018). However, these former public servants work alongside people who have experience in different parts of the private sector or who are entering the workplace for the first time. These teams bring together a diverse range of skills, ranging from in-depth technical expertise to foreign languages and cultural expertise.

Cyber intelligence teams tend to be divided between technical and strategic functions (Chismon and Ruks 2015). This is by necessity, as it is rare for a single person to have the education and experience to be both a malware reverse engineer and an expert on geopolitics. This division of expertise can also be seen in CyCon's separation of technical and policy tracks. As with CyCon's focus on collaboration, effective cyber intelligence functions are those that can sustain collaboration between people with different forms of expertise.

Commercial cyber intelligence exists in a range of forms. These include organic intelligence functions within companies; vendors that provide cyber intelligence products and services; and intelligence service functions within cloud service providers (Work 2023). These companies and other cyber intelligence actors, in turn, form heterogeneous intelligence networks, sharing intelligence and collaborating in a range of public and non-public forums, as part of a community that '[spans] public and private domains' (Work 2020, 279). Highlighting the heterogeneity of these arrangements, Kalkman and Wieskamp identify four types of cyber intelligence networks: centralized networks, business networks, operational networks, and local networks (Kalkman and Wieskamp 2019, 4). The role of large transnational technology companies in the sector means that these networks also transcend state boundaries.

Cyber intelligence networks are, in turn, part of the broader 'cyber ecosystem' (Ensor 2022). Bonfanti describes this ecosystem as the '(not-formalized) international cybersecurity community that consists of representatives from supranational institutions and agencies, domestic public bodies, private organizations, and academia' (Bonfanti 2018, 105). Cyber intelligence networks therefore bring together actors that 'differ in terms of their history, activities, governance structure, communication frequency, goal consensus, member commitment, and perceived results' (Kalkman and Wieskamp 2019, 4). This heterogeneity creates multiple challenges, including differences in interests; in attitudes towards classification, ethics, and legal responsibility; in working practices; and in the use of language (Kalkman and Wieskamp 2019; Miller 2010). From the level of the individual up to the ecosystem, cyber intelligence is defined by two characteristics – heterogeneity and the importance of collaboration between different actors.

3. TWO SCENARIOS FOR FUTURE COLLABORATION

As indicated in the quotation from Lt. Gen. Copinger-Symes at the beginning of this paper, militaries are recognizing that the private sector has important cyber intelligence capabilities and that making the best use of these capabilities requires a close working relationship. This section outlines two scenarios for the future of this civil-military cyber intelligence collaboration. Greater collaboration between militaries and the private sector is the goal in both scenarios. Where the two scenarios differ is in the approach taken to achieving this collaboration. This paper proposes a distinction between approaches that emphasize unity and those that emphasize coherence. In the former, collaboration is advanced by one actor that imposes forms of organization, concepts, and language on other actors in this space. In the latter, actors seek instead to manage differences between themselves in a process that relies on people (here termed 'translators') who can work across different organizations and enable such interaction.

A. The Difference Between Unity and Coherence

The difference between unity and coherence can be seen in NATO standardization processes. The goal of these processes is not to eliminate differences in the equipment and procedures of NATO militaries but to manage these differences to enable effective collaboration:

> Interoperability is not centrally about the elimination of difference among national militaries and the production of a monolithic transnational NATO military. Rather, it is about coordinating difference and making it manageable, organizing bodies and materials in ways that produce new capacities. (Dittmer 2017, 81)

Pursuing coherence rather than unity requires a degree of comfort with ambiguity and a recognition that it is possible to collaborate with actors whose interests intersect – but do not fully align – with our own. This comfort with ambiguity and emphasis on working with and through partners resembles, to some extent, the mindset of counterinsurgency or intelligence (Healey 2020). By contrast, an approach aimed at unity seeks to impose order and 'eliminate ambiguity' (Kramer, Butler, and Lotrionte 2017, 14).

The UK's Industry 100 (i100) scheme is an example of collaboration through 'coherence'. Under the i100 scheme, individuals from industry and academia can work with the National Cyber Security Centre (NCSC) on secondment. The NCSC is a 'unique construct' that has been described as the 'Switzerland of Cyber' – a place 'where competitors from the private sector come together with government staff,

on neutral territory' (W 2022). The impetus for the creation of i100 came not from the NCSC but from a senior business leader in the private sector who presented the proposal to the NCSC and offered to provide funding (Ashdown 2024, forthcoming). In a similar vein, the NCSC operates a workspace on the popular collaboration platform Slack for UK cyber defenders. That workspace was originally created as a grassroots platform for discussion by a business leader in the private sector before it was handed over to the NCSC to operate (Ashdown 2024). The organic development and transformation of these initiatives, as opposed to their creation and enforcement by the state, evinces an approach that aims at coherence rather than unity.

B. Scenario 1: Unity

In this scenario, over the next decade, Western governments will impose forms of organization on the private sector in a primarily top-down approach. The state adopts a coordinating role, creating the structures through which collaboration is permitted and required to take place. The emphasis is on the private sector coming to the public sector rather than vice versa. In this scenario:

- Collaboration between militaries and cyber intelligence providers is presented as overriding those companies' own interests. This approach is presented publicly as necessary to enable effective action on a matter of national security and defence.
- This coordinating role places greater emphasis on the state to be able to identify relevant actors, coordinate their activity in real time, and adapt to rapidly changing circumstances. The organized actors may also at times resist or struggle to fulfil their designated roles.
- As the coordinating actor, the state imposes conceptual frameworks and terminology on the private sector. Given the familiarity of militaries with establishing doctrine and training people at scale, these frameworks are likely to be primarily military in origin.
- The government and the military prioritize the development of in-house capability through strategies aimed at recruiting and retaining talent out of universities or from industry.
- Institutional arrangements are created that enable military personnel to dictate action to the private sector in times of war or crisis. More broadly, a top-down, unified approach establishes clear lines of responsibility between civil and military actors.

C. Scenario 2: Coherence

In this scenario, Western governments do not dictate the form of civil-military cyber collaboration but instead encourage the development of a heterogeneous ecosystem. Rather than central coordination, this approach emphasizes decentralized social

networks and emergent forms of organization. Government and military actors still operate top-down hierarchical structures for civil-military collaboration, but they coexist with alternative forms of organization. In this scenario:

- Collaboration takes place through the interlacing of multiple motives, including not only a desire to act in the interest of national security but also financial, reputational, and academic motives, among others.
- Personal relationships and trust are key to the working of this ecosystem. Much of the work falls on people who can act as 'translators' between different groups (see below).
- Without a single central authority dictating practices, frameworks, and terminology, a more complicated situation arises, with different groups adopting different approaches and language to tackle similar issues. Enabling cooperation between these groups is a key role of the translators described above.
- In this scenario, the movement of people from the public to the private sector is seen less as a retention challenge and more as an opportunity to seed the ecosystem with talent and to build personal connections into industry.
- The heterogeneity of commercial cyber intelligence is maintained in this scenario, in which collaboration takes place through surprising and novel organizational constructs and forums, often driven by industry participants rather than by the government or the military. However, this heterogeneity complicates efforts to clearly delineate roles and responsibilities between the state and private actors. Attempts to impose central coordination in times of crisis or conflict will partly rely on ad hoc cooperation rather than the activation of pre-existing lines of command.

The key differences between these two scenarios are set out in Table I below.

TABLE I: COMPARING SCENARIOS – UNITY AND COHERENCE

Unity	Coherence
National security drives collaboration	Multiple motives for collaboration
State plays key coordinating role	Social networks play key coordinating role
Adoption of standard frameworks	Translation between frameworks
Retention of trained personnel	Movement encouraged to develop ecosystem
Clarity over roles and responsibilities	Reliance on ad hoc cooperation

4. DISCUSSION

There are advantages and disadvantages to each aspect of the two scenarios set out above. The following discussion section will weigh some of these competing points, to underline why a coherence-focused approach is ultimately preferable for civil-military collaboration in cyber intelligence. Finally, it will highlight the tendency for advocates of the unity approach to adopt the language of coherence.

A. Cognitive Dispersion versus Productive Ambiguity

A key advantage of the unity approach is that by imposing one set of conceptual frameworks, it overcomes the challenges of collaboration between people with 'very different mental models, jargon, and methodological approaches' (Vogel et al. 2017, 173). The claim that unity in terminology is necessary for effective collaboration in the cyber domain has been advanced by senior academics and practitioners (Neal-Hopes 2011; Lin 2020). Neal-Hopes uses the biblical story of Babel – a city whose people were condemned to speak different languages – as an analogy to argue for the unification of terminology:

> Unfortunately, the language of cyber space is a contemporary city of Babel…. The lack of a common lexicon detailing cyber's roles is producing cognitive dispersion at a time when the efficient expansion and aggregation of cyber forces demands cohesion. (Neal-Hopes 2011, 39)

An advocate of the coherence approach would argue that there is an important difference between shared understanding and imposing a single organization's terminology. The former is necessary for effective collaboration; the latter is likely to be counterproductive. As an example of this dynamic, the term 'open source' can refer to unclassified information or to software with public source code. This is exactly the kind of term that produces misunderstandings in collaborations between intelligence practitioners and technical experts (Vogel et al. 2017). However, attempting to enforce unity in the use of the term would be unhelpful. What matters is that people understand what their collaborators mean when they use the term.

It is entirely possible for a military cyber defence unit to work with a civilian contractor without both having to adopt explicitly military terminology, such as 'kill chain' (Hutchins, Cloppert, and Amin 2011). This may even bring benefits. As Miller observes, 'The military and intelligence establishments use words and concepts that … often tend to be misconstrued or misinterpreted by those outside their circle' (Miller 2010, 696). Moreover, all use of language encodes implicit ideas that may prematurely circumscribe the possibilities of collaboration between actors (Slayton 2021; Branch 2020). Insisting that all activity in cyberspace can only be accurately

described through the language of warfare is likely to prove counterproductive if the goal is to encourage effective collaboration with the private sector (Slayton 2021; Gravell 1998). However, it also precludes the possibility of drawing on frameworks and terminology from a range of different areas, from both government and the private sector. Organizations capable of tolerating such ambiguity find themselves not with cognitive dispersion but rather with a broader range of conceptual tools with which to approach complex and unfamiliar problems.

B. The Need for Translators versus the Benefits of Centralization

Adopting a coherence approach therefore puts a premium on the work of people who can act as mediators – or 'translators' – between different organizations and cultures, helping those groups to engage productively with each other. Explaining differences between professional lexicons is one part of this role, but it also involves negotiating issues of difference and identity within organizations more broadly. Effective collaboration entails changes in ways of thinking and acting among all the parties involved. Translators are individuals who, through their interactions and movements across organizational boundaries, facilitate these productive transformations (Ashdown 2024).

The importance of translators is well understood, and they appear in a variety of guises across the literature. Harknett and Stever note approvingly that 'dual-hatted' officials who sit in different institutional contexts can provide 'greater synergy of perspective' (Harknett and Stever 2009, 7). Vogel et al. describe the importance within cross-organizational teams of 'human enablers' 'who can advise on how to work best collaboratively' (Vogel et al. 2017, 178). Trent emphasizes the importance of 'spanners and brokers' who can work across organizations, 'moving ideas and building new communities'; he emphasizes that spanners are 'people with (or willing to develop) first-hand experience in multiple domains' (Trent 2018, 120).

However, this is not an easy role to play, and a coherence approach places a substantial burden on these translators (Ashdown 2024). It is for this reason that the unity approach reflects a preference for collaboration to be centralized and managed at a higher level rather than depending on social networks and professional relationships. The goal is 'to transcend the dependence on individual personal/professional relationships which are, to an extent, transient' (Piazza, Vasudevan, and Carr 2023, 11). A centralized approach has the advantage of reducing the burden on the people whose personal relationships currently maintain networks of collaboration in public and private cyber intelligence. However, as will be argued below, the risk is that in centralizing control, much of the benefit of those informal, haphazard, but productive connections will be lost.

C. Tensions Between Motives for Collaboration

Collaboration on cyber intelligence makes the private sector an active participant in the cyber defence of military networks, in a way that Lt. Gen. Copinger-Symes argues differs from the traditional defence procurement relationship, 'where ultimately industry partners hand over a tank and then we operate the tank' (Martin 2023). Such relationships are an example of the kind of 'operational intimacy' between public and private actors that, it is argued, will be required to strengthen cyber security (Inglis and Krejsa 2022; Pell 2022; Landau 2022). However, such active engagement and proximity to operational military activity by private sector actors raises policy, legal, and ethical questions and challenges core ideas about the primacy of the state in matters of defence (Carr 2016; Miller 2010; Abrahamsen and Williams 2010).

The unity approach simplifies some of these questions – although it does not resolve them – by creating clear lines of control and responsibility for private actors. In this scenario, the reason private actors collaborate with the state on cyber defence is that they are mandated to do so in the national interest (Harknett and Stever 2009; Kramer, Butler, and Lotrionte 2017; Carr 2016). Some advocates of this approach have proposed creating nascent command structures for the private sector that can be activated in times of crisis or conflict (Kramer, Butler, and Lotrionte 2017). By contrast, in the coherence scenario it is much more challenging to see how different actors' interests are aligned and where they differ, and where the divisions between roles and responsibilities lie.

The challenge with the unity approach is that it risks casting the private sector as a passive intermediary rather than as a diverse set of actors with their own interests and motives. For example, Harknett and Stever's discussion of the benefits of greater public–private collaboration portrays non-state actors as impersonal bodies that are 'to be mobilized' and that must 'submit and participate' in the federal government's initiatives (Harknett and Stever 2009, 2, 10). They argue that public concern over privacy should be downplayed in favour of 'the more important issue of the general population's responsibility to government' (Harknett and Stever 2009, 2). Similar perspectives are visible in the claim that an effective 'whole of society' approach requires the private sector to 'adjust to a paradigm shift … where national security concerns must be elevated above corporate interests' (Williams 2023, 3). Such terminology treats other actors in the network as intermediaries that must be aligned with the state's interests. The goal is 'more subordination than alliance' (Healey 2020, 93). As noted above, these arguments address pressing questions about the organization of civil-military cooperation in the cyber domain. Yet the way they are articulated precludes further discussion. Rather than encouraging debate, they announce a conclusion that another actor has failed properly to internalize.

D. The Risk of Pursuing Unity in a Heterogenous, Collaborative Ecosystem

Crucially, even if we accept these arguments about national security, blunt statements of necessity may not be the best way to promote willing collaboration. Research suggests that cyber intelligence networks that emerged in a 'bottom-up' fashion are perceived by their members as producing better results (Kalkman and Wieskamp 2019, 11, 19–20). Similarly, an analysis of cyber intelligence and security collaboration in the US notes:

> The sharing of threat intelligence information based on current threat activity is … embraced in the private sector. This sharing is not bound by a common framework or lexicon and is driven by private cybersecurity companies. (Bronk and Conklin 2022, 161)

This is the fundamental point about cyber intelligence that argues for an approach aimed at coherence. Analysis of innovative technological ecosystems suggests that they develop through risk-taking, cooperation, and social networks rather than through 'hierarchical organizations, vertical information flow, and centralized decision-making' (Trent 2018, 117–18). With a capability as specialized, transnational, and fundamentally collaborative as cyber intelligence, any attempt to impose unity risks undermining the very capability it seeks to mobilize. In this vein, Healey and Korn argue that calls for the private sector to work under US government control 'may be counterproductive':

> The main goal of the coordination of all of these defensive efforts … is not unity of command centered on [the US Department of Defense], but unity of effort, unity of action, and loose coordination to keep independent groups working toward the same goal. (Healey and Korn 2019, 235)

Such an approach would mean that the Department of Defense would exercise less control and that defence would be less well coordinated – 'but this is a small loss to achieve better synchronization across all defense, in both the public and private sectors' (Healey and Korn 2019, 235). Preserving what makes cyber intelligence valuable requires an approach that can accept variation and ambiguity, one that seeks coherence rather than unity. Adopting such an approach will be challenging. It will require organizations to embrace productive ambiguity rather than fearing cognitive dispersion; to trust in individual social networks and translators as well as hierarchical structures; and to respect the diverse set of motives that might lead people and organizations to collaborate. However, the advantages of this approach outweigh these challenges.

5. CONCLUSION

The production and use of cyber intelligence, and the practice of cyber security more broadly, involves bringing together complex networks of people, processes, and technologies in a way that is fundamentally innovative and collaborative (Kalkman and Wieskamp 2019). This presents a challenge for NATO militaries, as they increasingly collaborate with the private sector in this area. Civil-military collaboration on cyber intelligence will play out in multiple ways simultaneously, through contractual arrangements, intelligence sharing platforms and forums, the movement of individuals between careers, and informal personal connections between individuals. NATO militaries will need to adapt to working in unfamiliar ways with an important new set of partners. In this regard, Healey and Korn call into question 'the default assumption of military cyber defenders that, to defend the Nation, they must take control of the assets themselves' (Healey and Korn 2019, 231). Overcoming that 'default assumption' – the desire to impose unity rather than live with ambiguity and difference – will be crucial for effective military collaboration with the private sector.

Highlighting the need for change, Chris Inglis and Harry Krejsa call for 'a new social contract for the digital age – one that meaningfully alters the relationship between public *and private sectors and proposes a new set of obligations for each'* (Inglis and Krejsa 2022, emphasis added). Such calls should always be viewed with some caution. The goal of coherence is a naturally appealing one, particularly in cyber security. The idea that digital technologies necessitate novel approaches, particularly ones that are perceived as more agile and decentralized, has been formative in understandings of the cyber domain (Arquilla and Ronfeldt 1993). The attractiveness of this idea means there is a risk that people will use the language of 'coherence' to describe approaches to managing civil-military collaboration that fall back on 'unifying' practices of centralization and control.

A foreseeable consequence of genuine collaboration will be a transformation of *all* the organizations involved, public and private. Vogel et al. argue that the key to effective collaboration between academia, industry, and the intelligence community is 'organizational innovation and adaption' (Vogel et al. 2017, 174). This process of transformation is something that militaries have at times opposed, as seen, for example, in resistance to lowering physical fitness requirements for personnel working in cyber roles (Lin 2020, 97). In his September 2023 interview, Lt. Gen. Copinger-Symes described the creation of the UK National Cyber Force – a novel organization bringing together Defence and the intelligence agencies in a way that has required changes for all participants. Lt. Gen. Copinger-Symes described this as a process of 'bending ourselves out of shape together' (Martin 2023). This image captures the essence of the transformative nature of collaboration – different entities

coming together, with a willingness to productively manage difference and, in doing so, to develop new capacities (Dittmer 2017).

Theorists of cyber conflict suggest we can learn from studying the development of concepts and operational art around earlier novel domains such as the air and the sub-surface (Neal-Hopes 2011; Hurley 2012; Perkovich and Levite 2017). However, the value of historical study may be as much in examining the process by which multiple actors collectively (but not always cooperatively) defined those new domains and their roles within them. In doing so, those actors redefined themselves – they 'bent themselves out of shape together'. It is precisely such a process of redefinition that will play out as militaries and private sector actors learn how to collaborate in the production, sharing, and use of cyber intelligence. The choice facing these organizations is whether to pursue coherence and bend, or cling to unity and break.

REFERENCES

Abrahamsen, Rita, and Michael C. Williams. 2010. *Security beyond the State: Private Security in International Politics*. Cambridge University Press.

Andrew, Christopher. 2018. *The Secret World: A History of Intelligence*. Penguin UK.

Arquilla, John, and David Ronfeldt. 1993. 'Cyberwar Is Coming!' *Comparative Strategy* 12 (2): 141–65.

Ashdown, Neil. 2024. 'Advocates of Collaboration: Assembling Cyber Intelligence in the UK'. PhD Thesis, Royal Holloway University of London.

Bellaby, Ross W. 2016. 'Justifying Cyber-Intelligence?' *Journal of Military Ethics* 15 (4): 299–319. https://doi.org/10.1080/15027570.2017.1284463.

Bonfanti, Matteo E. 2018. 'Cyber Intelligence: In Pursuit of a Better Understanding for an Emerging Practice'. *Cyber, Intelligence, and Security* 2 (1): 105–21.

Branch, Jordan. 2020. 'What's in a Name? Metaphors and Cybersecurity'. *International Organization*, September, 1–32. https://doi.org/10.1017/S002081832000051X.

Bronk, Chris, and Wm Arthur Conklin. 2022. 'Who's in Charge and How Does It Work? US Cybersecurity of Critical Infrastructure'. *Journal of Cyber Policy* 7 (2): 155–74. https://doi.org/10.1080/23738871.2022.2116346.

Carr, Madeline. 2016. 'Public–Private Partnerships in National Cyber-Security Strategies'. *International Affairs* 92 (1): 43–62. https://doi.org/10.1111/1468-2346.12504.

Chismon, David, and Martyn Ruks. 2015. *Threat Intelligence: Collecting, Analysing, Evaluating*. MWR InfoSecurity.

Dittmer, Jason. 2017. *Diplomatic Material: Affect, Assemblage, and Foreign Policy*. Duke University Press.

Ensor, Chris. 2022. 'NCSC View: Future Ecosystem Challenges'. NCSC. 1 November 2022. https://www.ncsc.gov.uk/search.

Gravell, William. 1998. ‘Some Observations Along the Road to National Information Power Symposium: International Information Infrastructure Protection and National Security’. *Duke Journal of Comparative & International Law* 9 (2): 401–26.

Guerrero-Saade, Juan Andrés. 2015. ‘The Ethics and Perils of APT Research: An Unexpected Transition into Intelligence Brokerage’. In *Proceedings of the 25th Virus Bulletin International Conference*. http://media.kaspersky.com/pdf/Guerrero-Saade-VB2015.pdf.

Harknett, Richard J., and James A. Stever. 2009. ‘The Cybersecurity Triad: Government, Private Sector Partners, and the Engaged Cybersecurity Citizen’. *Journal of Homeland Security and Emergency Management* 6 (1). https://doi.org/10.2202/1547-7355.1649.

Healey, Jason, and Erik B. Korn. 2019. ‘Defense Support to the Private Sector: New Concepts for the DoD’s National Cyber Defense Mission’. *The Cyber Defense Review*, 227–44.

Healey, Jason. 2020. ‘A Bizarre Pair: Counterinsurgency Lessons for Cyber Conflict’. *Parameters* 50 (3): 85–94.

Hurley, Matthew M. 2012. ‘For and from Cyberspace: Conceptualizing Cyber Intelligence, Surveillance, and Reconnaissance’. *Air & Space Power Journal* 26 (6): 12–33.

Hutchins, Eric M., Michael J. Cloppert, and Rohan M. Amin. 2011. ‘Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains’. *Leading Issues in Information Warfare & Security Research* 1 (1): 80.

Inglis, Chris, and Harry Krejsa. 2022. ‘The Cyber Social Contract’. *Foreign Affairs*, 21 February 2022. https://www.foreignaffairs.com/articles/united-states/2022-02-21/cyber-social-contract.

Kalkman, Jori Pascal, and Lotte Wieskamp. 2019. ‘Cyber Intelligence Networks: A Typology’. *The International Journal of Intelligence, Security, and Public Affairs* 21 (1): 4–24. https://doi.org/10.1080/23800992.2019.1598092.

Kramer, Franklin D., Robert J. Butler, and Catherine Lotrionte. 2017. ‘Cyber and Deterrence’. Atlantic Council. 3 January 2017. https://www.atlanticcouncil.org/in-depth-research-reports/report/cyber-and-deterrence/.

Landau, Susan. 2022. ‘Cyberwar in Ukraine: What You See Is Not What’s Really There’. Default. 30 September 2022. https://www.lawfaremedia.org/article/cyberwar-ukraine-what-you-see-not-whats-really-there.

Lin, Herbert. 2020. ‘Doctrinal Confusion and Cultural Dysfunction in DoD: Regarding Information Operations, Cyber Operations, and Related Concepts’. *The Cyber Defense Review* 5 (2): 89–108.

Lindsay, Jon R. 2020. ‘Cyber Conflict vs. Cyber Command: Hidden Dangers in the American Military Solution to a Large-Scale Intelligence Problem’. *Intelligence and National Security*, October 2020. http://www.tandfonline.com/doi/abs/10.1080/02684527.2020.1840746.

Martin, Alexander. 2023. ‘British Army General Says UK Now Conducting “Hunt Forward” Operations’. 25 September 2023. https://therecord.media/uk-hunt-forward-operations-lt-gen-tom-copinger-symes.

Mattern, Troy, John Felker, Randy Borum, and George Bamford. 2014. ‘Operational Levels of Cyber Intelligence’. *International Journal of Intelligence and CounterIntelligence* 27 (4): 702–19. https://doi.org/10.1080/08850607.2014.924811.

Miller, Bowman H. 2010. ‘Soldiers, Scholars, and Spies: Combining Smarts and Secrets’. *Armed Forces & Society* 36 (4): 695–715. https://doi.org/10.1177/0095327X10361667.

Neal-Hopes, Timothy. 2011. ‘“Preventing a Cyber Dresden”: How the Evolution of Air Power Can Guide the Evolution of Cyber Power’. Master’s thesis, School of Advanced Air and Space Studies. https://apps.dtic.mil/sti/pdfs/AD1019450.pdf.

Pell, Stephanie. 2022. ‘Private-Sector Cyber Defense in Armed Conflict’. *Lawfare*, 1 December 2022. https://www.lawfareblog.com/private-sector-cyber-defense-armed-conflict.

Perkovich, George, and Ariel Levite, eds. 2017. *Understanding Cyber Conflict: 14 Analogies*. Washington, DC: Georgetown University Press.

Petersen, Karen Lund, and Vibeke Schou Tjalve. 2018. ‘Intelligence Expertise in the Age of Information Sharing: Public–Private “Collection” and Its Challenges to Democratic Control and Accountability’. *Intelligence and National Security* 33 (1): 21–35. https://doi.org/10.1080/02684527.2017.1316956.

Piazza, Anna, Srinidhi Vasudevan, and Madeline Carr. 2023. ‘Cybersecurity in UK Universities: Mapping (or Managing) Threat Intelligence Sharing within the Higher Education Sector’. *Journal of Cybersecurity* 9 (1): tyad019. https://doi.org/10.1093/cybsec/tyad019.

Slayton, Rebecca. 2021. ‘What Is a Cyber Warrior? The Emergence of U.S. Military Cyber Expertise, 1967–2018’. *Texas National Security Review*, 11 January 2021. http://tnsr.org/2021/01/what-is-a-cyber-warrior-the-emergence-of-u-s-military-cyber-expertise-1967-2018/.

Smith, Brad. 2022. ‘Defending Ukraine: Early Lessons from the Cyber War’. Microsoft On the Issues. 22 June 2022. https://blogs.microsoft.com/on-the-issues/2022/06/22/defending-ukraine-early-lessons-from-the-cyber-war/.

Stout, Mark, and Michael Warner. 2018. ‘Intelligence Is as Intelligence Does’. *Intelligence and National Security* 33 (4): 517–26. https://doi.org/10.1080/02684527.2018.1452593.

Trent, Stoney. 2018. ‘Cultivating Technology Innovation for Cyberspace Operations’. *The Cyber Defense Review* 3 (3): 115–34.

Vogel, Kathleen M., Jessica Katz Jameson, Beverly B. Tyler, Sharon Joines, Brian M. Evans, and Hector Rendon. 2017. ‘The Importance of Organizational Innovation and Adaptation in Building Academic–Industry–Intelligence Collaboration: Observations from the Laboratory for Analytic Sciences’. *The International Journal of Intelligence, Security, and Public Affairs* 19 (3): 171–96. https://doi.org/10.1080/23800992.2017.1384676.

W, Ollie. 2022. ‘Inside Industry 100 - the on-Loan CTO’. NCSC.GOV.UK. 22 April 2022. https://www.ncsc.gov.uk/blog-post/inside-industry-100-the-on-loan-cto.

Warner, Michael. 2014. *The Rise and Fall of Intelligence: An International Security History*. Georgetown University Press. http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=766382&site=ehost-live.

Wiener, Craig. 2016. ‘Penetrate, Exploit, Disrupt, Destroy: The Rise of Computer Network Operations as a Major Military Innovation’. PhD thesis, George Mason University.

Williams, B. K. 2023. *Public-Private Bonhomie May Not Last Amid Strategic Competition*. Livermore, CA: Lawrence Livermore National Laboratory (LLNL).

Work, J. D. 2020. ‘Evaluating Commercial Cyber Intelligence Activity’. *International Journal of Intelligence and CounterIntelligence* 33 (2): 278–308. https://doi.org/10.1080/08850607.2019.1690877.

———. 2023. ‘Private Actors and the Intelligence Contest in Cyber Conflict’. In *Deter, Disrupt, or Deceive: Assessing Cyber Conflict as an Intelligence Contest*, edited by Robert Chesney and Max Smeets, 225–60. Georgetown University Press. https://books.google.co.uk/books?id=ZHmbEAAAQBAJ.

Zegart, Amy. 2020. ‘Intelligence Isn’t Just for Government Anymore’. *Foreign Affairs*, 5 November 2020. https://www.foreignaffairs.com/articles/united-states/2020-11-02/intelligence-isnt-just-government-anymore.

CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)

Innovations in International Cyber Support: Comparing Approaches and Mechanisms for Cyber Capability Support

Joseph Jarnecki
Research Fellow
Royal United Services Institute
London, United Kingdom
jarneckijoseph@gmail.com

Abstract: This paper proposes the concept of international cyber capability support (CCS) to describe a policy area covering the direct provision of cyber security products and services, including the deployment of rapid response teams, with immediate operational impacts intended to advance short-to-medium term objectives. This paper examines illustrative case studies of national cyber crises demonstrating the need for CCS deployment, emerging approaches to providing CCS, and the considerations which should inform how CCS is operationalized.

A recent uptick in large-scale national cyber incidents has demonstrated a clear need for international capabilities to support crisis scenarios. Responses to date have been ad hoc and have exploited crisis conditions to short-circuit normally slow decision-making processes when providing international cyber support. As these incidents become business as usual, formal mechanisms to provide rapid capability responses are being developed. This paper discusses these points, drawing on high-level case studies of Ukraine and Costa Rica.

Capability support outside of crises has also seen a marked increase, such as through personnel deployments as part of Hunt Forward Operations. As countries and international organizations look to establish CCS mechanisms, they need to consider their strategic objectives, implementation, thresholds for deployment and withdrawal, and the remit of their activities. Observing and learning from existing approaches is essential.

Challenges for actors operationalizing CCS mechanisms include aligning activities with partners, creating an enabling and legitimating environment, and monitoring, measuring, and assessing initiatives. From these, this report recommends that actors create a strategic value-case, consider carefully how to integrate multiple stakeholders into CCS mechanisms, and take a comprehensive approach to international cyber support.

Keywords: *international cyber support, cyber capability, cyber capacity building, national cyber crisis*

1. INTRODUCTION

There has been a worrying uptick in national, large-scale cyber incidents in recent years. Montenegro, Costa Rica, Vanuatu, Ukraine, and Albania have all experienced significant effects from cyber attacks; government systems have been crippled and the delivery of critical services delayed or denied.[1] Cyber crisis response, undertaken by countries and big tech companies, has attempted to mitigate the extent and impact of these incidents. To date, this support has been delivered in an agile and ad hoc manner, with minimal formal mechanisms in place. France, the United States, and Microsoft, for example, have reactively deployed teams to several national incidents.[2]

As actors establish formal mechanisms to deliver national cyber crisis responses, they face various challenges, trade-offs, and choices. Who should be providing support and where? What should support entail? How should it be organized? Is support a security, diplomatic, development, or humanitarian matter? And what are the thresholds for deployment and withdrawal? It is also necessary to consider how the provision of capability is operationalized outside of crises, building a cohesive approach to international cyber capability support (CCS).

This paper argues that international CCS mechanisms show promise but have been slow to get off the ground. Moreover, it asserts that emerging frameworks by governments and international organizations should make more use of lessons from recent incidents, particularly those involving multiple implementing actors. Even the most well-resourced actors have limited capabilities to scale up and extend the provision of support, and it remains uncertain whether companies will continue to assist as they have, for example, in Ukraine.

1 Sydney J. Freedberg Jr., 'State Dept Wants "Cyber Assistance Fund" to Aid Allies and Partners Against Hackers', *Breaking Defense*, 10 April 2023; Taylor Grossman, *Cyber Rapid Response Teams Structure, Organization, and Use Cases*, (Zürich: Center for Security Studies ETH Zurich, 2023), 30.

2 Mubariz Zaman, 'Montenegro Thanks France for Assistance Following Cyberattacks', *Diplomatic Insight*, 29 August 2022.

The paper first provides brief case studies of existing ad hoc international cyber support in national cyber crises, focusing on Ukraine and Costa Rica. It then highlights emerging national and international mechanisms for CCS. Finally, the paper identifies significant implications for governments when creating and operationalizing CCS mechanisms, and it outlines priority considerations in a policy-relevant format.

What Is Cyber Capability Support?

This paper uses the term 'cyber capability support' (CCS) to describe a policy area involving the direct provision of products, services, or other cyber security solutions, including rapid crisis response, which have immediate operational impacts that advance short-to-medium-term objectives. For example, CCS includes purchasing licenses, deploying/contracting incident responders, and providing/purchasing actionable cyber threat intelligence (CTI). Activities such as building a security operations centre, conducting a national cyber maturity review, or assessing existing legislation do not fall within this remit; these are more aligned with cyber capacity building (CCB).

CCB is a well-established policy area within international cyber support.[3] Attempts have been made to apply a CCB lens to understand international support responding to national cyber crises and to bring the direct provision of capabilities within CCB frameworks. The desire to not 'reinvent the wheel' is laudable, but CCB frameworks and terminology should not be expected to fit every situation. Policies that aim to create endogenous capacity within recipients should be distinguished from those that provide or deploy capabilities to recipients. The former consists of sustainable, long-term activities, whereas the latter involves operational, responsive, and dynamic support, including in crisis scenarios. Further distinguishing features between CCS and CCB are outlined in Table I.

TABLE I: CCS AND CCB DISTINGUISHING FEATURES

Features	Cyber Capability Support (CCS)	Cyber Capacity Building (CCB)
Timeline	• Rapid deployment, implying dynamic procurement solutions • Typically short-term deployments (<1 year)	• Open competition procurement • Usually mid-to-long-term deployment (>1 year) with limited exceptions
Intended outcome	Achieving targeted and immediate operational objectives to strengthen partner's cyber security resilience and protection in the short-to-medium term, including denying and disrupting adversary activities. Capabilities are provided or purchased to be used by the recipient or are delivered by an implementor.	Creating and supporting endogenous recipient capacity to internally anticipate and respond to cyber risks and threats, including through targeting tactical and strategic outcomes such as improved population cyber hygiene.

3 *See, e.g.*, the 200+ members and partners of the Global Forum for Cyber Expertise, a CCB-focused international organization. 'Members & Partners', GFCE, accessed 4 March 2024, https://thegfce.org/member-and-partner/.

Examples of activities	• Provision of cyber security services, e.g., incident response, remediation • Provision of cyber security products, e.g., firewalls, attack-surface management	• National assessments • Strategy development • Awareness campaigns • Training and education • Limited provision of technical products and services
Examples of policy instruments	• Rapid Response Teams and Mechanisms • Hunt Forward Operations	• Cybersecurity Capacity Maturity Model • National Cyber Risk Assessment • Cyber Defense Exercise with Recurrence
Withdrawal threshold	End of operation, though this is inconsistently defined	End of project or programme

Author-generated

Other attempts to systematize CCS as a policy area have described it as 'cyber defense assistance',[4] 'deployed cyber defence',[5] and 'cybersecurity support deployments'.[6] This paper eschews each of these concepts in an attempt to avoid military language and to posit CCS as a separate but complementary partner to CCB, though it accepts the need to further discuss and align understandings and terminology.

This paper anticipates the critique that CCB is a sufficient frame for all international cyber support and that CCS is not needed. While the sentiment is understandable, it is valuable to discuss and reassess approaches to policy. As international cyber support receives more attention and funding, it is important to demarcate policy areas by type of activity or intended outcome. While CCB and CCS project activities may overlap in places, Table I outlines a sensible set of criteria to divide those building capacities and those providing capabilities. Moreover, the intended outcome of CCB, which is to develop and sustainably build the recipient's own capacities, is not identical to CCS outcomes, which involve providing or purchasing assistance to achieve operational objectives that immediately improve the recipient's cyber security. These differences can be acknowledged without undermining a shared overarching strategic objective: improving and supporting partners' cyber resilience.

This paper focuses primarily on civilian components of CCS and draws exclusively from open-source material. A further study of emerging CCS mechanisms would benefit from primary data-gathering with relevant policymakers.

4 Rattray, Brown, and Moore, 'The Cyber Defense Assistance Imperative Lessons'.

5 Nick Beecroft and Toby Gilmore, 'The Advantages of "Hunt Forward" Extend Beyond the Hunt', *BAE Systems Digital Intelligence*, 2023.

6 Julia Schuetze and Eglė Daukšienė, 'Cybersecurity Support Deployments: An Emerging Cooperative Approach', Stiftung Neue Verantwortung, 15 June 2023, https://www.stiftung-nv.de/en/publication/cybersecurity-support-deployments.

2. TWO CASE STUDIES: UKRAINE AND COSTA RICA

This section addresses two recent case studies of rapid international cyber crisis response to demonstrate the increasing salience of and need for these activities, as well as their use of ad hoc processes.

A. Case Study: Sustained Russian Cyber Campaign against Ukraine

Russian cyber operations against Ukraine began scaling up before the invasion in February 2022 and have continued throughout the war. Reports from that period indicate a shift from sophisticated and long-term cyber operations to intelligence gathering and less sophisticated destructive tactics.[7] Significant effects have been observed on Ukrainian critical infrastructure, such as the 2024 attack on the telecom company Kyivstar.[8]

Before the invasion, several actors were conducting CCB in Ukraine. These included the European Union, the US, Estonia, France, the United Kingdom, and Germany, focused on areas such as cybercrime, cyber hygiene, and awareness building.[9] As war became more likely, some actors undertook CCS, deploying targeted services to improve Ukrainian systems resilience. Public information on these activities is limited; however, the US Cyber Command's (USCYBERCOM) Hunt Forward Operation (HFO) has been disclosed publicly. An HFO involves deployed personnel hunting for threats on partner networks alongside local counterparts.[10] The mission to Ukraine (December 2021 to February 2022), which was praised by a senior Ukrainian cyber security official, included the discovery of ninety malware samples.[11]

Once the war began, cyber support became one part of a broader assistance to Ukraine. Limited information is available on support to the Ukrainian Defence Ministry and armed forces; a notable exception is the IT Coalition, which is part of the Ramstein format.[12]

7 *See, e.g.*, 'Cyber Conflict in the Russia-Ukraine War', Carnegie Endowment, accessed 3 January 2024, https://carnegieendowment.org/programs/technology/cyberconflictintherussiaukrainewar/; Google TAG and Mandiant, 'Fog of War', Google, February 2023, https://blog.google/threat-analysis-group/fog-of-war-how-the-ukraine-conflict-transformed-the-cyber-threat-landscape/.

8 Tom Balmforth, 'Exclusive: Russian Hackers Were Inside Ukraine Telecoms Giant for Months', *Reuters*, 5 January 2024, https://www.reuters.com/world/europe/russian-hackers-were-inside-ukraine-telecoms-giant-months-cyber-spy-chief-2024-01-04/.

9 For CBB project details, *see* Cybil Portal, 'Projects', Cybil, accessed 4 January 2024, https://cybilportal.org/projects-advanced/?_sft_country=ukraine&_sfm_status_project=Finished.

10 US Cyber Command Public Affairs, 'CYBER 101: Hunt Forward Operations', US Cyber Command, 15 November 2022, https://www.cybercom.mil/Media/News/Article/3218642/cyber-101-hunt-forward-operations/.

11 Dina Temple-Raston et al., 'Exclusive: Ukraine Says Joint Mission with US Derailed Moscow's Cyberattacks', *Record*, 18 October 2023, https://therecord.media/ukraine-hunt-forward-teams-us-cyber-command.

12 European Pravda, 'Ramstein Format Meeting: 10 IT Coalition Countries Sign 6-Year Cooperation Agreement', *Ukrainska Pravda*, 14 February 2024, https://www.pravda.com.ua/eng/news/2024/02/14/7441891/.

There is more open-source information on non-military CCS activities by countries and private companies. Among foreign governments, the UK moved first to establish the Ukraine Cyber Programme (UCP) shortly after the invasion and has since welcomed funding from other international donors to expand activities.[13] The UCP has utilized cyber security providers to supply incident response (IR), DDoS (distributed denial-of-service) protection, firewalls, and forensic capabilities.[14] Germany has sent cyber security hardware as humanitarian aid to Ukraine's energy sector.[15] Microsoft, Google, and other private companies have provided various licenses, tools, and technical assistance.[16] Other international actors have also conducted CCS, with the general approach being ad hoc cyber assistance to Ukraine.[17]

Efforts to coordinate CCS to Ukraine began in the private sector with the Cyber Defence Assistance Collaborative (CDAC).[18] A volunteer group of cyber security and technology companies established in March 2022, the CDAC minimized the risk of duplication and streamlined assistance.[19] Governments have been slower to establish similar structures, with the Tallinn Mechanism announced in December 2023 and the Ramstein format IT Coalition formalized in February 2024.[20] The Tallinn Mechanism's three chronological lines of effort – 'short (Support), medium (Build) and long-term (Sustain)' – imply that CCS-type activities are covered in addition to CCB, as does its commitment to 'maintain and strengthen' Ukrainian cyber resilience.[21] If this interpretation is correct, the Tallinn Mechanism coordinates a hybrid of CCS and CCB activities.

This paper considers CCS to Ukraine to include rapidly deployed activities taken to directly provide cyber security products or services with immediate operational impacts on advance short-to-medium-term objectives. For example, USCYBERCOM's HFO deployed personnel to provide threat hunting services, and the UCP purchases products and services for Ukrainian systems resilience. By contrast, CCB initiatives

13 Prime Minister's Office, 'UK to Give Ukraine Major Boost to Mount Counteroffensive', GOV.UK, 18 June 2023, https://www.gov.uk/government/news/uk-to-give-ukraine-major-boost-to-mount-counteroffensive.

14 FCDO, 'UK Boosts Ukraine's Cyber Defences with £6 Million Support Package', GOV.UK, 1 November 2022, https://www.gov.uk/government/news/uk-boosts-ukraines-cyber-defences-with-6-million-support-package.

15 Cybil Portal, 'Supporting Ukraine's Cybersecurity Agency with Hardware', *Cybil*, accessed 4 January 2024, https://cybilportal.org/projects/supporting-to-ukraines-cybersecurity-agency-with-hardware/.

16 Nick Beecroft, 'Evaluating the International Support to Ukrainian Cyber Defense', Carnegie Endowment, 3 November 2022, https://carnegieendowment.org/2022/11/03/evaluating-international-support-to-ukrainian-cyber-defense-pub-88322.

17 'Tallinn Mechanism', Estonian Ministry of Foreign Affairs, https://www.vm.ee/en/international-law-cyber-diplomacy/cyber-diplomacy/tallinn-mechanism.

18 Greg Rattray, Jeff Brown, and Robert T. Moore, 'The Cyber Defense Assistance Imperative Lessons from Ukraine', Aspen Institute, February 2023, https://creativecommons.org/licenses/ by-nc/4.0/.

19 Rattray, Brown, and Moore, 'The Cyber Defense Assistance Imperative Lessons'.

20 Other parties to the mechanism include Canada, France, Germany, the UK, Netherlands, Poland, Sweden, Ukraine, and the US.

21 'Tallinn Mechanism', Government of Canada, last modified 19 December 2023, https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/peace_security-paix_securite/tallinn-mechanism-mecanisme-tallinn.aspx?lang=eng.

such as USAID's 'Cybersecurity for Critical Infrastructure' project, while significant and substantial, have long-term, primarily strategic objectives to develop Ukraine's own capacities. Grouping both sets of activities – CCB and CCS – within a single policy framework that necessitates similar approaches to processes such as funding and procurement makes it difficult to respond appropriately to donor objectives and recipient needs.

B. Case Study: Ransomware Attacks on Costa Rica

'We are determined to overthrow the government by means of a cyber attack …' – Conti, a ransomware group[22]

In spring 2022, Costa Rica experienced a series of cyber attacks which led the president to declare a national emergency and announce that the country was 'at war'.[23] Beginning on 17 April, the ransomware group Conti launched attacks in rapid succession, impacting twenty-nine government institutions.[24] These abated in early May, but new attacks, now by the Hive ransomware group, started on 31 May, targeting the Costa Rican Social Security Fund.[25]

The campaigns by Conti and Hive disrupted the delivery of critical services. Ministry of Finance digital systems to declare taxes and customs were shut down. So were some Social Security Fund services, which affected an estimated 4,871 medical appointments in the initial twenty-four hours.[26] Response costs for the government were over US$24 million as of June 2022, and economic losses from disruptions to trade have been estimated at US$38 million per day.[27]

In one of the first actions taken in response to the attacks, the government of Costa Rica asked Spain, the US, and Israel for advice and support.[28] Spain, which had a pre-existing agreement with Costa Rica, sent 100,000 licenses for a counter-ransomware tool and deployed a government technical team. Israel, which had cemented bilateral cyber relations with a memorandum of understanding (MoU), provided CTI.[29] The US deployed an FBI technical team and offered a US$10 million reward for information

22 Jonathan Grief, 'Ransomware Gang Threatens to "Overthrow" New Costa Rica Government, Raises Demand to $20 Million', *Record*, 16 May 2022, https://therecord.media/ransomware-gang-threatens-to-overthrow-new-costa-rica-government-raises-demand-to-20-million.

23 'President Rodrigo Chaves says Costa Rica is at war with Conti hackers', BBC News, 18 May 2022, https://www.bbc.com/news/technology-61323402.

24 Eugenia Lostri and Georgia Wood, 'The Role of International Assistance in Cyber Incident Response', *Lawfare*, 30 March 2023, https://www.lawfaremedia.org/article/role-international-assistance-cyber-incident-response.

25 Jonathan Grieg, 'Costa Rican Social Security Fund Hit with Ransomware Attack', *Record*, 31 May 2022, https://therecord.media/costa-rican-social-security-fund-hit-with-ransomware-attack.

26 Andrea More, 'CCSS report afectación de 4.871 usarios en 80 establecimientos de salud, tras hackeo a sistemas informáticos', *Delfino*, 1 June 2022, https://delfino.cr/2022/06/ccss-reporto-afectacion-de-4-871-usuarios-en-80-establecimientos-de-salud-tras-hackeo-a-sistemas-informaticos.

27 Lostri and Wood, 'The Role of International Assistance'.

28 Ibid.

29 Ibid.

on Conti's leadership.[30] Microsoft and Cisco supplied free tools, and Microsoft provided unspecified additional technical assistance.[31] This report argues that these support efforts fall under CCS.

To the author's knowledge, no actor providing CCS to Costa Rica had specific policies in place to conduct that kind of activity. Moreover, there was no prior decision-making process to determine that – along with the private sector – the US, Spain, and Israel were appropriate and legitimate responders. Nor were there thresholds, at least publicly, in place to determine what triggered a request for assistance. Presumably, once they were engaged, supporting actors deconflicted their activities, but this is not certain. Equally unclear was how they determined when to withdraw support – the US continued to launch new initiatives into 2023, though these focused on building capacity.[32] Did the US deliberately transition its activities from providing capability to building capacity in Costa Rica?

CCS to Costa Rica was welcomed and had an impact. President Chaves has stated he has '25 million reasons to be grateful' for the cyber security support.[33] A high-profile attack has also motivated other Latin American countries, with many subsequently creating national cyber security strategies – from twelve in 2020 to over twenty by 2024.[34]

While individual national preparations are welcome, measures to formalize international CCS are moving more slowly. A warning from Conti, released during its attacks on Costa Rica, should encourage these processes: 'The Costa Rica scenario is a beta version of a global cyber attack on an entire country'.[35]

30 Ned Price, 'Reward Offers for Information to Bring Conti Ransomware Varient Co-Conspirators to Justice', US Department of State, Press Statement, 6 May 2022, https://www.state.gov/reward-offers-for-information-to-bring-conti-ransomware-variant-co-conspirators-to-justice/.

31 Lostri and Wood, 'The Role of International Assistance'.

32 'United States Announces $25 Million to Strengthen Costa Rica's Cybersecurity', US Embassy in Costa Rica, 29 March 2023, https://cr.usembassy.gov/united-states-announces-25-million-to-strengthen-costa-ricas-cybersecurity/.

33 Luke O'Grady, 'Event Recap: A Conversation with Rodrigo Chaves Robles, President of Costa Rica', Center for Cybersecurity Policy and Law, 6 November 2023, https://www.centerforcybersecuritypolicy.org/insights-and-research/event-recap-a-conversation-with-rodrigo-chaves-robles-president-of-costa-rica.

34 Cecilia Tornaghi, 'The Dramatic Cyberattack that Put Latin America on Alert', *Americas Quarterly*, 25 July 2023.

35 VenariX (@_venarix_), '#Conti's latest update on the cyberattack…', X, 20 April 2022, https://twitter.com/_venarix_/status/1516569937418113025.

3. EMERGING APPROACHES

This section outlines and considers existing, emerging, or proposed mechanisms and programmes to coordinate and conduct CCS. Private-sector-led initiatives such as the CDAC are excluded in the interest of brevity. This section supports the argument that existing and emerging CCS mechanisms are making good progress but that there are diverse issues to consider and best practices to integrate.

Table II provides an overview of CCS mechanisms and programmes for which there is open-source information.

TABLE II: OVERVIEW OF CCS MECHANISMS AND PROGRAMMES

Entity	Mechanism	Description
European Union (EU)	Cyber Reserve[36]	• Proposal under the EU Cyber Solidarity Act • Private IR services deployable at the request of EU members or organizations • Response to significant or large-scale cyber security incidents • Funding for whole proposed Act (including other provisions) is €1.1 billion
US State Department	Cyberspace, Digital connectivity, and related Technology (CDT)[37]	• State Department fund, including CCS such as emergency assistance capacities • Deployed at the discretion of the secretary of state • Created under the 2023–2024 Department of State Authorization Act • US$150 million for five-year period from 1 October 2023
Ukraine Defence Contact Group (UDCG)	IT Coalition[38]	• Ten-country initiative within the Ramstein Format coordinating defence support to Ukraine • Established with six-year commitment to deliver secure and resilient IT infrastructure for Ukrainian defence forces • Funding unclear, with some individual members announcing contributions, e.g., €10 million each from both Luxembourg and the Netherlands
UK Foreign Commonwealth & Development Office (FCDO)	Ukraine Cyber Programme (UCP)[39]	• Direct programme to provide CCS to Ukraine, supporting networks against Russian attacks • Launched February–March 2022 • Procurement of private providers by FCDO • £7.1 million programme expenditure with a further up to £25 million of multi-country funding committed from June 2023 – potential £9 million from allies

36 36 "The EU Cyber Solidarity Act', European Commission, updated 6 March 2024, https://digital-strategy.ec.europa.eu/en/policies/cyber-solidarity.

37 "Text - S.2043 - 118th Congress (2023-2024): Department of State Authorization Act of 2023', Congress.gov, 22 August 2023, https://www.congress.gov/bill/118th-congress/senate-bill/2043/text/is.

38 'Increased Support from Partners, Capabilities Coalition and IT Coalition - Outcomes of the 15th Ramstein Meeting', Ministry of Defence of Ukraine, Government Portal, 19 September 2023, https://www.kmu.gov.ua/en/news/bilshe-pidtrymky-vid-partneriv-capabilities-coalition-it-koalitsiia-pidsumky-15-oi-zustrichi-u-formati-ramshtain; Viktoria Stepanenko, New Air Defense Coalition and Military Aid Agreed at Latest Ramstein Meeting', *Kyiv Post*, 24 November 2023, https://www.kyivpost.com/post/24566; 'Netherlands Joins IT Coalition to Support Ukraine and Contributes Over $10 mn', Army Recognition, 29 January 2024, https://armyrecognition.com/ukraine_-_russia_conflict_war_2022/netherlands_joins_it_coalition_to_support_ukraine_and_contributes_over_$_10_mn.html?utm_content=cmp-true.

39 Prime Minister's Office, 'UK to give Ukraine Major Boost'.

USCYBERCOM	HFO[40]	• Rooted in 2018 Department of Defense Cyber Strategy doctrine of defend forward and persistent engagement • Involving the deployment of USCYBERCOM operators to hunt for threats alongside host nation counterparts • Over 50 deployments to 24 countries[41]
EU Permanent Structured Cooperation (PESCO)	Cyber Rapid Response Teams (CRRT)[42]	• Multi-country/pooled IR capability; nine members as of September 2023[43] • Launched in February 2018 • Provision of emergency response, confidence-building, and training to partners who request it • Teams composed of government experts • Deployments to Ukraine, Mozambique, and Moldova • Operations funded jointly by providers and recipients
Counter Ransomware Initiative (CRI)	CRI Incident Response (CRI-IR)[44]	• Fifty country members committed, under 2023 joint statement, to assist in IR if government or lifeline sectors are hit by ransomware • Few details available
Australia Department of Foreign Affairs & Trade (DFAT)	Regional Cyber Crisis Response Team (R-CCRT)[45]	• Mechanism to provide CCS • Limited to Pacific and Southeast Asian countries experiencing severe cyber incidents • Coordinated by Australia's Cyber Ambassador within DFAT • Funding around A$26–A$43 million
Foreign ministries: US, Canada, Denmark, Estonia, France, Germany, the Netherlands, Poland, Sweden, UK	Tallinn Mechanism[46]	• Coordinates and facilitates 10 countries' civilian international cyber support (CCS & CCB) to Ukraine • Launched in December 2023 • NGO and private sector involvement stated but not detailed • Presumably no funding distribution function, just deconfliction and coordination among members
North Atlantic Treaty Organization (NATO)	Virtual Cyber Incident Response Capability (VCISC)	• Country members provide assistance for post-incident mitigation • Launched at Vilnius Summit July 2023 • Few operational details available

Author-generated

A. Organization

Understanding how mechanisms are organized is key to understanding their proliferation, as well as their intended and actual abilities. Australia's R-CCRT and the US's CDT and HFOs are national mechanisms, whereas the EU's CRRT and

40 US Cyber Command Public Affairs, 'CYBER 101: Hunt Forward Operations', US Cyber Command, 15 November 2022, https://www.cybercom.mil/Media/News/Article/3218642/cyber-101-hunt-forward-operations/.

41 Patty Nieberg, '"Hunt Forward" Cyber Teams Have Deployed to 24 Countries, Including Ukraine', *Task & Purpose*, 28 September 2023, https://taskandpurpose.com/news/cyber-command-security-hunt-forward/.

42 PESCO, 'Cyber Rapid Response Teams and Mutual Assistance in Cyber Security (CRRT)', PESCO Projects, accessed 11 January 2024, https://www.pesco.europa.eu/project/cyber-rapid-response-teams-and-mutual-assistance-in-cyber-security/.

43 Belgium, Denmark, Estonia, Croatia, Lithuania, Poland, Netherlands, Romania, and Slovenia.

44 'International Counter Ransomware Initiative 2023 Joint Statement', White House, 1 November 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/11/01/international-counter-ransomware-initiative-2023-joint-statement/.

45 Australian Government, *Australian Cyber Security Strategy 2023-2030* (Canberra: Australian Government Department of Home Affairs, 2023).

46 "Tallinn Mechanism', Estonian Ministry of Foreign Affairs, 21 December 2023, https://www.vm.ee/en/international-law-cyber-diplomacy/cyber-diplomacy/tallinn-mechanism.

Cyber Reserve, NATO's VCISC, the UDCG's IT Coalition, the Tallinn Mechanism, and the CRI-IR are multinational. The UK's UCP began as a national programme but ostensibly became a multi-country-funded mechanism (see Table II).

1) Multinational

The EU's CRRT is a limited member mechanism coordinated by permanent and rotating co-chairs; one co-chair, Lithuania, is permanent. The CRRT's response teams are supposed to consist of experts from multiple members, though how often this is actually achieved is uncertain.[47] Both the IT Coalition and the Tallinn Mechanism are ten-country initiatives, and the latter states that it also involves tech companies and NGOs. The former is headed by Estonia and Luxembourg, and the latter has a front office in Kyiv run by Estonia and a back office in Poland.

The full scope of the EU Cyber Reserve remains unclear, largely because it is in development, though initial indications are that it will procure assistance, funded by the EU, from a pool of private incident responders. While details of its projected funding are unclear, it is likely over €100 million.[48]

The UK's UCP was launched as a national initiative, which, in June 2023, welcomed additional funding from international partners, thus becoming multinational.

The CRI-IR and NATO's VCISC are at present too opaque to draw conclusions about their organization.

2) National

The US CDT, which operates alongside USCYBERCOM HFOs, deployments by the FBI, and CCB delivered by USAID, is part of an increasingly comprehensive approach the country is taking to international cyber support. Acknowledging this, the authorization of funds for the CDT is contingent upon a 'review of emergency assistance capacity' by the secretary of state within a year.[49] Australia's R-CCRT aligns with the country's stance that cyber response can be a humanitarian activity and builds on the regional focus of its existing CCB.[50] The significant funding that the US and Australian governments have allocated to new CCS mechanisms indicates their increasing focus on the policy area (see Table II).

B. Remit

Policymakers coordinating mechanisms and programmes in Table II have sought to elaborate clear remits to avoid mission creep.

47 Grossman, 'Cyber Rapid Response Teams'.

48 'The EU Cyber Solidarity Act', European Commission, updated 6 March 2024, https://digital-strategy.ec.europa.eu/en/policies/cyber-solidarity.

49 Congress.gov, 'Text - S.2043'.

50 Cybil Portal, 'Projects', Cybil, accessed 10 January 2024, https://cybilportal.org/projects-advanced/?_sft_funder=australia.

The listed initiatives' remits can be grouped across membership, geography, and priority/opportunity criteria (see Table III). The CRI-IR, NATO's VCISC, and the EU's Cyber Reserve provide membership-based support, which may be mediated through a multinational body. The UCP, the Tallinn Mechanism, and the IT Coalition, as well as Australia's R-CCRT, have geographical focuses – Ukraine for the first three and the Pacific Islands and Southeast Asia for the last one. Finally, the US CDT and HFOs have looser remits, each determined by government priorities. For example, the CDT states that its activities are intended to 'advance a stable and secure cyberspace' and 'support and reinforce democratic values and human rights'.[51] The CRRT ostensibly focuses on EU member states and organizations, but in practice it has, for example, supported Mozambique and Moldova.

TABLE III: OVERVIEW OF MECHANISM REMITS

Remits		
Membership	Geography	Priority/Opportunity
• CRI-IR • VCISC • Cyber Reserve	• UCP • Tallinn Mechanism • IT Coalition • R-CCRT	• CDT • HFO • CRRT

Author-generated

C. Thresholds

Factors that shape the thresholds for deployment of a given mechanism include the scale of incidents, types of attack, and political decision-making.

A significant, large-scale crisis or emergency or a severe incident is the threshold for response for NATO's VCISC, the EU's Cyber Reserve and CRRT, Australia's R-CCRT, and the CRI-IR. However, none of these mechanisms define precisely what this means. Given that they are programmes, not reproducible mechanisms, the IT Coalition, UCP, and the Tallinn Mechanism do not provide thresholds for further deployment, but each stemmed from Russia's large-scale cyberattacks on Ukraine.

The threshold for CRI-IR activation is limited to ransomware attacks that hit government or lifeline sectors.

US HFOs and the CDT have no publicly disclosed thresholds for deployment. Instead, they activate at the discretion of USCYBERCOM and the secretary of state, respectively.

51 Congress.gov, 'Text - S.2043'.

Thresholds for withdrawal are less elaborated and highly sensitive; publicly defining a limit to support may encourage adversaries or undermine relationships with partners. As such, all mechanisms but the IT Coalition have avoided setting hard limits. The IT Coalition has set a six-year horizon on its activities; this might be a response to domestic pressures, though no reason has been stated publicly.

D. Private Sector Involvement

The mechanisms in Table II have considered how to include private companies; a high-level overview is provided in Table IV.

One approach, adopted by the US's CDT and the UK's UCP, is to use private sector companies as implementing partners. Under this approach, companies operationalize CCS. The EU Cyber Reserve has similarly indicated it would use private-sector implementation and would maintain a list of trusted providers. The Tallinn Mechanism goes further by referring to tech companies and NGOs in donor countries as mechanism participants, though it does not clarify what this entails. To the author's knowledge, US HFOs have never integrated private sector provision. According to Taylor Grossman, the EU's CRRT had intended to do so but has been hampered by classification and liability issues.[52]

Australia's R-CCRT commits to drawing on industry experience. The author's assumption is that this involves contracting private sector implementation, though this has not been clarified publicly. The capabilities which the IT Coalition and NATO's VCISC commit to providing – IT infrastructure and national mitigation – imply private sector delivery, though, again, this has not been confirmed. Lastly, the CRI has sought to incorporate the private sector in its wider initiatives; however, their role in the CRI-IR commitment has not been discussed publicly.

TABLE IV: ROLE OF PRIVATE SECTOR

Private Sector Involvement			
Implementation Partners	Mechanism Partners	Not Involved	Unclear
• CDT • UCP	• Tallinn Mechanism • Cyber Reserve	• HFO • CRRT	• VCISC • R-CCRT • CRI-IR • IT Coalition

Author-generated

52 Grossman, 'Cyber Rapid Response Teams', 18.

4. OPERATIONALIZING CCS

This section outlines thematic issues for CCS provision and priority considerations for mechanisms. These are summarized by single words or phrases to promote their adoption by policymakers.

This section supports the argument that emerging CCS mechanisms are promising and that key considerations are coming to the fore. Actors intending to create or improve their CCS offering should learn from the existing efforts of like-minded partners.

A. Alignment

Deconfliction. Points of duplication exist among mechanisms outlined in Table II. Papua New Guinea, for example, is a CRI member and is covered by Australia's R-CCRT. If Papua New Guinea suffers a severe incident, who would respond? As more mechanisms emerge, especially those without geographical limitations, there is a risk of further duplication. While efforts to deconflict mechanisms seem desirable, having overlapping coverage could mitigate the risk of overloading one mechanism. Managing these overlaps, however, will be difficult. Actors want to respond to the most severe and strategically significant incidents; thus, they might be incentivized to compete to provide capability in some cases but under-supply in others. This is further complicated by the involvement of big tech in CCS; for example, Microsoft has responded to national incidents in Costa Rica, Ukraine, Albania, and elsewhere.[53] The provision of CCS by diverse, multi-stakeholder actors creates a need to understand incentives and to conduct regular communication, coordination, and deconfliction. Efforts to coordinate diverse multi-stakeholder activities will inevitably encounter significant difficulties, as demonstrated by CCB, but this should not dissuade actors from making the attempt.

In addition to managing duplication among stakeholders, actors operationalizing CCS need to consider how to structure their activities around ongoing CCB. This paper advocates that CCS covers activities to provide capability, especially in anticipation of, during, and immediately after significant incidents. At any stage of CCS intervention, there could be previous, ongoing, or planned CCB. While CCS and CCB have distinct intended outcomes, they should not be conducted in isolation and should be joined up where possible. Actors offering broad international cyber support face a challenge in taking a comprehensive approach and ensuring that CCS and CCB activities are complementary.

53 Brad Smith, 'Defending Ukraine: Early Lessons from the Cyber War', Microsoft, 22 June 2022, https://blogs.microsoft.com/on-the-issues/2022/06/22/defending-ukraine-early-lessons-from-the-cyber-war/; Microsoft Threat Intelligence, 'Microsoft Investigates Iranian Attacks Against the Albanian Government', Microsoft, 8 September 2022, https://www.microsoft.com/en-us/security/blog/2022/09/08/microsoft-investigates-iranian-attacks-against-the-albanian-government/.

Cohesion. Actors conducting CCS should consider how to work with likeminded partners to ensure their efforts are cohesive and complementary. Multi-member CCS mechanisms pool resources, encourage economies of scale, and can streamline processes such as information sharing and requests for assistance. They also require collaborative objective setting; for instance, the CRI-IR has the strategic objective of mitigating the effects of ransomware. Setting and sticking to these objectives could prove difficult, as participants hold differing views. For instance, some participants argue for lower thresholds to provide CCS, while others prioritize attribution over remediation, and assembling joint teams can prove intractable, as gaps in trust prevent information sharing. These considerations for multi-member mechanisms are reproduced to some extent within the governments that run national mechanisms. Achieving cohesion is easier said than done and is it not necessarily always possible or desirable.

B. Enabling Environment

What can be done. Mechanisms, whether national or multinational, require that actors or groups split up certain competencies across people, process, and technology.

- Implementing personnel can be direct employees of funders, as with USCYBERCOM's HFOs, or contracted. Required personnel are not just technical individuals deploying tools and services; CCS projects also need strong project and stakeholder management. To sustain a standing mechanism, donors may require a secretariat, as the Tallinn Mechanism shows (see Section 3.A).
- Clear and targeted processes are important for the success of CCS mechanisms. This is especially the case for rapid response, where internal processes could activate teams that are on standby 24/7. Alternatively, if private sector deployment is leveraged, a rapid response capability could be serviced by a retainer arrangement, where a company is paid a fee to ensure that responders are on hand whenever necessary.
- Technological resources to provide CCS can be held or created by actors directly or acquired from industry, such as the UCP purchasing forensic capabilities.[54]

Given that mechanisms may need to address incidents of increasing complexity, and multiple incidents simultaneously, they need to be scalable. This paper suggests that a sufficient capacity to scale up CCS is unlikely to be achieved outside the most well-resourced countries but is more feasible in a multi-country mechanism. A more realistic approach, however, may be to design a multi-stakeholder mechanism that integrates private sector delivery, though this raises questions about whether companies can legitimately provide this support, whether principals and shareholders

54 FCDO, 'UK boosts Ukraine's Cyber Defences'.

believe it is worthwhile, and whether companies share values and objectives with states and international organizations. On the other hand, are CCS mechanisms possible without the private sector? While national in-house CCS mechanisms may present fewer challenges in setting strategic direction, integrating multi-stakeholder implementing partners is likely essential to achieving comprehensive coverage.

Mechanisms which leverage private sector delivery, especially for rapid response, depend on effective funding and procurement processes. Actors should consider whether their existing processes are appropriate – for example, whether sufficient funding is eligible for capability provision and whether procurement processes accommodate cyber security providers with minimal experience in other international support activities. Hopefully, donors and researchers will conduct further analyses of funding and procurement issues.

What should be done. Actors operationalizing CCS need to consider their legitimacy to act. National and international law is decisive in determining legitimacy. For example, some countries' constitutions prohibit or limit foreign security assistance, and international law presents considerations related to requests for assistance.[55] Actors providing CCS in limited geographical or country contexts could account for these issues by pre-agreeing MoUs with partners, though these require extended and complex negotiations. For less targeted mechanisms, substantial effort is required to anticipate legal issues such as data-sharing, classifications, procurement, and funding rules.

C. Monitoring and Measurement

Justification. The perceived success of CCS mechanisms depends on how they are monitored, measured, and evaluated. Donors need to understand the value-case of activities and whether their implementation provides a sufficient return on investment. None of the actors running the mechanisms outlined in Table II have released information on these considerations, but similar thinking on CCB indicates it is a consideration.[56] Part of assessing the value-case involves actors deciding strategic priorities for CCS, such as humanitarian, security, commercial, or influence priorities. For example, NATO's VCISC is focused on national mitigation among allies; Australia's R-CCRT supports regional partners, presumably in the interest of building influence; and the US CDT integrates some commercial priorities. Furthermore, those determining value should expect questions about moral hazard: does the creation or provision of CCS disincentivize national cyber security preparation to avoid large-

55 Net Politics, 'How Japan's Pacifist Constitution Shapes its Approach to Cyberspace', Council on Foreign Relations, 23 May 2018, https://www.cfr.org/blog/how-japans-pacifist-constitution-shapes-its-approach-cyberspace; Louise Marie-Hurel, 'Decoding Emerging Threats: Ransomware and the Prevention of Future Cyber Crises', RUSI, 11 September 2023, https://rusi.org/explore-our-research/publications/conference-reports/decoding-emerging-threats-ransomware-and-prevention-future-cyber-crises.

56 Faisal Hameed et al., 'Analysing Trends and Success Factors of International Cybersecurity Capacity-Building Initiatives', in *Twelfth International Conference on Emerging Security Information, Systems and Technologies*, 2018.

scale incidents? From this question, some actors may decide to impose conditions on assistance, though this could impact relationships with partners. Worse still, will the existence of these mechanisms incentivize adversaries to conduct attacks triggering responses? Could the resource strain of providing CCS divert resources better spent elsewhere?

Challenge and response. Monitoring and measuring policy mechanisms is a persistent challenge. To monitor mechanisms, implementers and funders need oversight and access to data on the impacts of implementation as well as an assessment framework. While monitoring and measuring is not easy, this paper asserts that it is possible to do so for CCS while acknowledging that some indicators (e.g., number of attacks) are clearer than others (e.g., deterring adversaries).

The data-rich nature of many CCS activities provides a valuable opportunity to monitor knowable or calculable impacts. For example, providing a cloud-based malware analysis tool could involve implementers receiving information on recipient tool usage as part of product improvement cycles. Similarly, IR services will involve implementers collecting data such as malware samples. If relevant calculable information was captured and submitted as part of activity assessment frameworks, CCS projects could be effectively measured and evaluated. Gathering and exfiltrating data, however, may be unacceptable to recipients who are concerned about unauthorized access by malicious actors.

Assessing CCS impacts also relies on subjective or qualitative judgements based on smaller or more opaque data sets. For example, assessing CCS activities intended to deny or deter adversaries relies on an understanding of adversary perceptions and reactions. The need to make this judgement is nothing new; actors developing CCS mechanisms should consider assessment approaches from existing foreign, defence, and security policy.

D. Considerations for CCS Mechanisms

Establishing and maintaining CCS mechanisms is complex. This paper argues that the most urgent points for actors to consider are creating a measurable strategic value-case, determining multi-stakeholder involvement in CCS, and creating a comprehensive approach to international cyber support.

A measurable strategic value-case. As with other international support activities, CCS does not have a single strategic value-case across donors. Currently, mechanisms' strategic objectives include development and humanitarian support, security, and diplomacy. Strategic objectives shape actors' provision of CCS, as they influence factors such as potential partners, eligible funding, responsible internal agencies,

ability to scale, and thresholds for provision and withdrawal. Actors' ability to measure and assess CCS activities with reference to these strategic objectives will be decisive in determining whether mechanisms are maintained in the medium term. While CCS mechanisms now seem to be proliferating rapidly, there is no guarantee that this will continue or that they will be beneficial.

Multi-stakeholderism. This paper has focused on CCS provided by countries and international organizations and has largely considered the private sector as an operational delivery partner. The CDAC in Ukraine and the provision of CCS-type activities in multiple national cyber incidents by companies such as Microsoft, however, demonstrate that private sector actors are independent strategic CCS players.[57] In this context, national and international CCS mechanisms must consider how, when, and in what way they engage with the private sector. The Tallinn Mechanism, for example, has ostensibly integrated the private sector, but the nature of this involvement is unclear. Does it go as far as CCB initiatives such as the Global Forum for Cyber Expertise, which have full private sector members?[58] The nature and extent of private sector involvement in CCS mechanisms will be decisive in shaping their priorities, capacities, and abilities to scale. This paper recommends that actors looking to establish or improve a CCS mechanism put significant effort into considering private sector involvement.

A comprehensive approach to international cyber support. CCS mechanisms are largely being developed by actors who are already engaged in some kind of international cyber support activities. For some actors these mechanisms are entirely novel, for others they are based on previously ad hoc processes or are derived from other areas of policy, such as humanitarian support or military assistance. While this report welcomes greater attention to CCS-like functions, it strongly recommends ensuring that a focus on CCS, and particularly emergency provision, does not crowd out other international cyber support. As detailed above, CCS should not come at the expense of CCB – long-term, sustainable interventions to build the recipient's internal capacities. Committing resources to responsive capability provision seems to address more urgent needs, but it may not be the most efficient way to address foundational challenges. Ultimately, neither CCB nor CCS can cover all facets of international cyber support. Actors should instead leverage diverse tools across a spectrum of international cyber support activities which are prioritised and deployed based on their strategic objectives.

57 Beecroft, 'Evaluating the International Support to Ukrainian Cyber Defense'.
58 'Members & Partners', GFCE, accessed 10 January 2024, https://thegfce.org/member-and-partner/.

5. CONCLUSION

Large-scale, national cyber incidents and events have been a fixture of recent years. International responses to these have been significant and, in most cases, have been conducted on an ad hoc basis. The creation of new CCS mechanisms begins to provide more clarity on how, where, and with whom actors intend to provide support.

This paper has argued that multiple national and international actors have begun creating and scaling CCS mechanisms and that many of these emerging initiatives show promise. It has analysed existing and proposed mechanisms to identify factors that shape the provision of CCS and has advocated that these and other identified lessons be considered by actors operationalizing CCS. Furthermore, it has argued that outlining a clear strategic objective, considering multi-stakeholder collaboration, and creating a comprehensive approach to international cyber support are key to effective CCS.

To put these arguments forward, this paper has proposed CCS as a policy area and has drawn a distinction between it and CCB within broader international cyber support. While this paper acknowledges that such a reconceptualization requires further discussion, it maintains that distinct policy instruments are useful to meet separate outcomes and impact objectives.

Section 2 of this paper outlined case studies of national cyber incidents in Ukraine and Costa Rica and the ad hoc response from actors operationalizing CCS. Section 3 then analysed emerging, existing, and proposed CCS mechanisms to identify points which affect their organization, remit, thresholds, activities, and private sector involvement. Lastly, Section 4 outlined considerations for operationalizing CCS across alignment, enabling environments, monitoring, and measurement; it also proposed urgent points for actors to consider.

Further research on this topic should examine how the strategic value-case of CCS mechanisms affects their implementation and the factors that determine the participation of multiple stakeholders. As part of efforts to reevaluate international cyber support, policy-focused research creating a typology of activities would be invaluable. Furthermore, research on CCS funding and procurement mechanisms will help ensure the efficiency of emerging mechanisms. Finally, comparative in-depth analysis should be conducted on CCS in various large-scale national cyber incidents to identify tactical and operational best practices.

ACKNOWLEDGEMENTS

I am grateful to the reviewers of this paper for their patience and the significant time and effort they have expended in providing comments and suggestions.

I am also appreciative of insights received from friends and colleagues, including within the RUSI cyber team – specifically, the shrewd recommendations of Pia Hüsch, Hugh Oberlander and Conrad Prince.