Published by CCDCOE

2025

17th International Conference on Cyber Conflict: The Next Step

C. Kwan, N. Gratzer,
K. Podiņš, M. Tolppa (Eds.)

CYCON

IEEE
Advancing Technology
for Humanity

2025
17th INTERNATIONAL CONFERENCE ON CYBER CONFLICT: THE NEXT STEP

Copyright © 2025 by CCDCOE Publications. All rights reserved.

IEEE Catalog Number:	CFP2526N-PRT
IEEE Xplore Publication:	CFP2526N-ART
ISBN (print):	978-9916-9789-8-6
ISBN (pdf):	978-9916-9789-9-3

COPYRIGHT AND REPRINT PERMISSIONS

No part of this publication may be reprinted, reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the NATO Cooperative Cyber Defence Centre of Excellence (publications@ccdcoe.org).

This restriction does not apply to making digital or hard copies of this publication for internal use within NATO, or for personal or educational use when for non-profit or non-commercial purposes, providing that copies bear this notice and a full citation on the first page as follows:

2025 17th International Conference on Cyber Conflict:
The Next Step
C. Kwan, N. Gratzer, K. Podiņš, M. Tolppa (Eds.)
2025 © CCDCOE Publications

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

LEGAL NOTICE: This publication contains the opinions of the respective authors only. They do not necessarily reflect the policy or the opinion of NATO CCDCOE, NATO, or any agency or government. NATO CCDCOE may not be held responsible for any loss or harm arising from the use of information contained in this book and is not responsible for the content of external sources, including external websites referenced in this publication.

NATO COOPERATIVE CYBER DEFENCE CENTRE OF EXCELLENCE

The NATO Cooperative Cyber Defence Centre of Excellence (NATO CCDCOE) is the leading dedicated hub for NATO allies and like-minded nations to jointly raise their cyber defence capabilities. The heart of the Centre is a diverse array of international experts from military, governmental, academic, and industrial backgrounds, currently representing 39 member nations from across the globe.

The Centre provides valuable expertise on cyber defence across strategic, legal, operational, and technical realms. It conducts research, delivers training and exercises, and develops doctrines, standards, and concepts to support and strengthen collective cyber resilience.

The NATO CCDCOE focuses on strengthening national cyber capabilities in five key areas: conducting cyberspace operations within a common framework, integrating cyber considerations into joint and multi-domain operations, enabling multinational cyber operations, coordinating military-civilian cyber activities, and fostering public–private partnerships in cyber defence.

Among its flagship activities are the cyber exercises Locked Shields and Crossed Swords, as well as the annual International Conference on Cyber Conflict (CyCon).

Locked Shields is the largest and most complex international live-fire cyber resilience exercise in the world. Each year, cyber professionals participate in this exercise to hone their ability to defend national IT systems and critical infrastructure during real-time, simulated cyberattacks. The exercise features realistic scenarios and cutting-edge technologies, encompassing a full-spectrum cyber incident that challenges participants' technical, legal, strategic, and communication responses.

Crossed Swords focuses on training cyber specialists to execute full-spectrum offensive cyber operations in a simulated crisis environment. The exercise also supports military command elements in practising command and control of offensive cyberspace capabilities, contributing to a more integrated and responsive cyber force.

CyCon, hosted by NATO CCDCOE since 2009, has become a major multidisciplinary platform for discussing the legal, technical, policy, strategic, and military aspects of cyber conflict. This unique event gathers prominent experts and decision-makers from the global cyber defence community, featuring over 100 speakers and attracting more than 600 attendees from government, military, industry, and academia. CyCon is accompanied by the proceedings, a collection of cutting-edge research discussed at the conference.

At this year's CyCon, the NATO CCDCOE is proud to launch two new publications. In collaboration with the University of Exeter, the Ministry of Foreign Affairs of Estonia, and the Ministry of Foreign Affairs of Japan, the NATO CCDCOE will launch *The Handbook on Developing a National Position on International Law and Cyber Activities: A Practical Guide for States*. This Handbook offers a practical and structured approach for States to develop or to review a national position, helping to foster greater legal clarity, predictability, and stability in cyberspace. By outlining existing practices, shared challenges, and strategic considerations, it offers a key resource to governments, legal practitioners, and policymakers navigating the application of international law in the cyber context.

Also debuting is the *Cyber Commander's Handbook 2*, which bridges the gap from the strategic to the operational perspectives of cyberspace. Intended as a practical guide, the Handbook has been developed to support commanders and decision-makers in understanding, integrating, and employing cyber capabilities.

As cyber threats grow in complexity and scale, the NATO CCDCOE continues to play a vital role in helping NATO and its partners maintain the initiative and adapt to the rapidly evolving threat landscape.

As a NATO-accredited Centre of Excellence, the NATO CCDCOE is not part of the NATO command structure.

CYCON 2025 SPONSORS

PLATINUM SPONSORS

Google Cloud

paloalto®
NETWORKS

DIAMOND SPONSORS

aws

FORTINET®

Microsoft

TREND MICRO™

GOLD SPONSORS

BLACK HILLS
Information Security

SEALINGTECH®
a PARSONS Company

Silobreaker

SILVER SPONSORS

CROWDSTRIKE

TECHNICAL SPONSOR

IEEE®
Advancing Technology
for Humanity

TABLE OF CONTENTS

FOREWORD

Another year passes and tensions continue to increase, potentiated by the continued confluence of geopolitics, technology, economy and society. This is reflected in the theme of the 17th International Conference on Cyber Conflict (CyCon), ‘The Next Step’. We need to look ahead. The lines between defensive and offensive cyber operations, civilian law enforcement and military activity, as well as between peacetime, crisis and conflict, are more blurred than ever. The range of actors is unprecedentedly diverse, and their intentions differ greatly. Our adversaries are active 24/7, and we must be so too.

The question posed in our Call for Papers was: How do politicians and decision-makers, industry, lawyers, and technological pioneers adapt when the rules of the game are constantly evolving?

The CyCon 2025 Programme Committee is proud to present 14 papers that answer this question from legal, strategy and policy, and technical angles.

In the legal track, **Giulia Pavesi** and **Andrea Alberti** examine the legal and policy interplay between the space and cyber domains in NATO’s collective defence framework, and highlight the fragmentation in the Alliance’s integration and coordination between these two interconnected areas.

Continuing the international law theme, **Gwendolyn Strasberg** and **Andom Gherezghiher** propose a prohibition on acts of aggression as a common legal foundation for collective countermeasures, a four-factor test for their deployment, and a clear, replicable threshold for their implementation. **Anna Blechová** examines the legal challenges involved in protecting critical infrastructure such as subsea cables and satellite infrastructure.

Lisandra Novo considers the risks of continued state engagement in the UN Convention on Cybercrime processes, including the issue of widening an already broad mandate. Examining another aspect of cybercrime, **Tsvetelina van Benthem** and **Roxana Radu** explore the interaction between domestic measures taken by states to counter threats from ransomware and their obligations under international law.

In the strategy and policy track, **Roxana Radu** continues the ransomware theme, presenting an overview of global trends in ransomware mitigation, highlighting the need to improve government coordination and reinforce public–private partnerships.

The blurring of traditional roles, such as civilian or military and public or private sectors, in terms of their contributions during conflict continues to be demonstrated by the Russia–Ukraine war. The role of Big Tech in military and civil defence is examined in a case study of the war by **Clara Cotroneo** and **Sarah Leonard**. The Russia–Ukraine war also saw the first call for voluntary civilian engagement in cyber conflict, sparking debate on the role of organised groups of volunteers acting upon state direction. **Gabrielle Joni Verreault** proposes an ethical framework to address challenges arising from voluntary civilian engagement in cyber and hybrid conflicts.

The unintended and malicious application of dual-use products and services has been an ongoing concern, and a concrete example is demonstrated by **Volodymyr Styran**, who analyses the dependency on Western cloud and IT infrastructure of Russian mobile applications used against Ukraine. Continuing the theme of dual-use products, **Ausma Bernot, M. Arif Khan, Khurram Shahzad, Mert Karakaya**, and **Conor Healy** demonstrate the weaknesses of legislative regulation of vulnerabilities in China-made Internet of Things surveillance cameras.

In the technical track, artificial intelligence (AI) is being put to good use. **Siam Shibly Antar, Philippe Charland, Steven H. H. Ding**, and **Benjamin C. M. Fung** demonstrate its utility in code-level rule generation for vulnerability patch verification in military software systems. **Allard Dijk, Roland Meier, Cosimo Melella, Mauno Pihelgas, Risto Vaarandi**, and **Vincent Lenders** consider the benefits of generative AI through large language models for Blue Team automation in cyber exercises such as Locked Shields, the NATO CCDCOE's flagship cyber defence exercise. **Silvio Russo, Michele Colajanni**, and **Claudio Zanasi** propose a novel approach to strategic dynamic deception using generative AI combined with other technologies to improve proactive cybersecurity.

Last but not least, **Michael Felux, Benoit Figuet, Vincent Lenders, Raphael Monstein**, and **Martin Strohmeier** present a new tool exploring automatic dependence surveillance-broadcast data to identify global navigation satellite systems (GNSS) jamming hotspots and the operational and safety implications of such activities.

In accordance with CyCon tradition and Institute of Electrical and Electronic Engineers (IEEE) procedures, all papers published in these proceedings have been subject to double-blind peer review. We are grateful to the CyCon 2025 Academic Review Committee for taking the time from their busy and full days to review and provide comprehensive, constructive feedback to authors and assist the Programme Committee in the final selection of papers in this volume. In this context, we extend our gratitude once again to the IEEE and its Estonian section for their continued support and technical sponsorship of the proceedings and of CyCon 2025 as a whole.

Finally, the Editors would like to thank Jaanika Rannu for her logistical support in the production of these proceedings. Special mentions go to Lt. Col. Nuno Rodrigues for his Easychair skills and for supporting the abstract selection process, and to Dr Claire Kwan for her work on the strategy track.

Academic Review Committee Members for CyCon 2025

- LCdr Dr Bernt Åkesson, NATO CCDCOE, Estonia
- Siim Alatalu, International Centre for Defence and Security (ICDS), Estonia
- Janos Barbi, NATO CCDCOE, Estonia
- Dr Bernhards 'BB' Blumbergs, CERT.LV, Latvia
- Philippe Charland, Defence Research and Development Canada, Valcartier Research Centre, Canada
- Yongkuk Cho, NATO CCDCOE, Estonia
- Dr Sean Costigan, George C. Marshall Center, United States
- Maj. John Dall, NATO CCDCOE, Estonia
- Prof. Thibault Debatty, Royal Military Academy, Belgium
- LCdr Erdi Donmez, NATO CCDCOE, Estonia
- Dr Andrew C. Dwyer, Royal Holloway, University of London, United Kingdom
- Dr Amy Ertan, NATO International Staff, Belgium
- Dr Kenneth Geers, GSK, United States
- Keir Giles, Conflict Studies Research Centre, United Kingdom
- Cmdr Davide Giovanelli, Italian Navy, Italy
- Nathalie Gratzer, NATO CCDCOE, Estonia
- Shota Gvineria, Baltic Defence College, Estonia
- Prof. Kimmo Halunen, University of Oulu and National Defence University, Finland
- Dr Jakub Harašta, Masaryk University, Czech Republic
- Dr Trey Herr, American University School of International Service, United States
- Otakar Horák, NATO CCDCOE, Estonia
- Dr Pia Hüsch, Royal United Service Institute for Defence and Security Studies, United Kingdom
- Dr Gabriel Jakobson, United States
- Taťána Jančárková, National Cyber and Information Security Agency (NÚKIB), Czech Republic
- Aleksi Kajander, NATO CCDCOE, Estonia
- Dr Ágnes Kasper, NATO CCDCOE, Estonia

- Prof. Sokratis Katsikas, Norwegian University of Science and Technology, Norway
- Erik Kursetgjerde, NATO CCDCOE, Estonia
- Dr Claire Kwan, NATO CCDCOE, Estonia
- Dr Kristi Land, Ministry of Foreign Affairs, Estonia
- Lt. Col. (ret.) Franz Lantenhammer, Germany
- Prof. Martin Libicki, US Naval Academy, United States
- Liina Lumiste, University of Tartu, Estonia
- Commander Michael McCarthy, Office of the JAG, Canadian Armed Forces, Canada
- Prof. Dr Olaf Maennel, The University of Adelaide, Australia
- Dr Matti K. Mantere, Starship Technologies, Estonia
- Dr Roland Meier, Swiss Federal Office for Defence Procurement armasuisse, Switzerland
- Stefano Mele, Gianni & Origoni Law Firm, Italy
- Prof. Marko Milanovic, University of Reading, United Kingdom
- Dr Tal Mimran, Hebrew University of Jerusalem, Israel
- Tomáš Minárik, National Cyber and Information Security Agency (NÚKIB), Czech Republic
- Dr Dóra Molnár, National University of Public Service, Hungary
- Dr Jose Nazario, Google, United States
- Gry-Mona Nordli, Norwegian Armed Forces, Norway
- Maj. Erwin Orye, Defence Forces, Belgium
- Dr Anna-Maria Osula, Ministry of Foreign Affairs, Estonia
- Dr Magdalena Pacholska, Asser Institute, The Netherlands
- Dr Piroska Páll-Orosz, Ministry of Defence, Hungary
- Prof. Constantinos Patsakis, University of Piraeus, Greece
- Piret Pernik, NATO CCDCOE, Estonia
- Dr Mauno Pihelgas, Estonia
- Col. Dr Peter Pijpers, Faculty of Military Science, Ministry of Defence, The Netherlands
- Col. MMMag, DDDr. Karl Platzer, Austrian Armed Forces, Austria
- Prof. JUDr Radim Polčák, Masaryk University, Czech Republic
- Col. Graham Price, Australian Cyber Command, Australia
- Prof. Michael Raska, S. Rajaratnam School of International Studies (RSIS), Singapore
- Lt. Col. Nuno Rodrigues, NATO CCDCOE, Estonia
- Prof. Marco Roscini, University of Westminster, United Kingdom
- Kurt Sanger, Cybersecurity and Data Protection, Buchanan, Ingersoll & Rooney PC, United States
- Adv. Prof. Annita Larissa Sciacovelli, University of Bari Aldo Moro, Italy

- 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, Swiss Federal Office for Defence Procurement armasuisse, Switzerland
- Prof. Dan Svantesson, Bond University, Australia
- Maria Tolppa, NATO CCDCOE, Estonia
- Dr Jens Tölle, GFP, Germany
- Kristel Urke, Ministry of Defence, Estonia
- Dr Risto Vaarandi, Tallinn University of Technology, Estonia
- Ann Väljataga, NATO CCDCOE, Estonia
- Dr René Värk, Ministry of Foreign Affairs, 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
- Tyron C. Wangard, NATO CCDCOE, Estonia
- Prof. Sean Watts, United States Military Academy at West Point, United States
- Dr Laurin Weissinger, Tufts University, United States
- Cmdr Mike Widmann, NATO Maritime Command / US Navy, United States
- Ingrid Winther, Norwegian Armed Forces, Norway
- Lt. Col. Nick Wobma, NATO CCDCOE, Estonia
- Jan Wünsche, Swedish Armed Forces, Sweden
- Danielle Yeow, Centre for International Law, National University of Singapore, Singapore
- Philippe Zotz, Luxembourg Armed Forces, Luxembourg

CyCon 2025 Programme Committee

- Dr Claire Kwan, Chair
- CDR Jack Shis, Vice-Chair
- Nathalie Gratzer, Track Chair (Strategy)
- Kārlis Podiņš, Track Chair (Technology)
- Maria Tolppa, Track Chair (Law)

CyCon 2025: The Next Step
17th International Conference on Cyber Conflict
C. Kwan, N. Gratzer, K. Podiņš, M. Tolppa (Eds.)
2025 © NATO CCDCOE Publications, Tallinn

Permission to make digital or hard copies of this publication for internal use within NATO and for personal or educational use when for non-profit or non-commercial purposes is granted provided that copies bear this notice and a full citation on the first page. Any other reproduction or transmission requires prior written permission by NATO CCDCOE.

NATO at a Cross-Road Between Space and Cyber Threats: A Legal, Policy and Operational Assessment of the Way Forward

Giulia Pavesi
Research Fellow
European Space Policy Institute
giulia.pavesi@espi.or.at

Andrea Alberti
Junior Research Fellow
European Space Policy Institute
andrea.alberti@espi.or.at

Abstract: This paper explores the critical interplay between the space and cyber domains and its implications for NATO's collective defence framework in the context of 21st-century security dynamics. The increasing integration of space and digital technologies has not only transformed the geopolitical landscape but also introduced complex vulnerabilities, as evidenced by the rising frequency and sophistication of cyberattacks targeting space systems. Despite NATO's recognition of cyberspace (2014) and space (2019) as operational domains, the Alliance's approach remains fragmented, with limited integration and coordination between these two interconnected areas. This analysis will assess NATO's evolving policies, highlighting the earlier institutionalization of cyberspace as a core collective defence priority compared to space, which only achieved similar recognition in 2021. Nevertheless, NATO's framework for space remains cautious, constrained by its dependence on Member States' capabilities and the dual-use nature of space infrastructure. This study identifies significant gaps in legal frameworks, political consensus and operational coordination, particularly in establishing thresholds for invoking collective self-defence in response to cyberattacks on space systems. Likewise, from a legal perspective, this paper examines the challenges of applying international law to cyber threats in space, focusing on issues such as defining thresholds for the activation of Article 5. From a political perspective, it underscores the complexities of achieving consensus among Member States and maintaining NATO's credibility in addressing cyber threats to space assets. Operationally, it reveals deficiencies in command structures, technological asymmetries and the lack of cross-domain integration. Recommendations include the establishment of clearer thresholds for activating collective self-defence, cross-domain simulation exercises and standardized security practices to enhance NATO's resilience and adaptability in addressing the threats posed by the increasingly interconnected space-cyber environment.

Keywords: *cyberspace interdependence, Article 5 threshold, legal framework evaluation, operational coordination*

1. INTRODUCTION

In the increasingly interconnected context of the 21st century, the relationship between the space and cyber domains emerges as a crucial element in the dynamics of the international political and strategic arena. The progressive and increasingly widespread interdependence between space and digital technologies has radically transformed the way States interact and pursue their interests, giving rise to new opportunities, but also new challenges and pitfalls.[1] Despite these developments, the national and supranational regulatory systems for both cybersecurity and space security remain fragmented and not necessarily convergent, including within NATO.[2] In its strategic concept approved in Madrid in June 2022, the Alliance pointed out the risks from cyber threats to both orbital and ground components of space systems. These could range from the disruption of business continuity in space to material degradation or total incapacitation of the system, potentially triggering the activation of Article 5, on a case-by-case determination.[3]

The increased threat surface of the Alliance is not mere fiction. There are currently roughly 13,300 space objects in low Earth orbit (LEO), while another 480 are in medium Earth orbit (MEO) and geostationary orbit (GEO). Of these, 12,500 are commercial satellites (10,800 in LEO and 1,700 in MEO or GEO) and 985 are reportedly military satellites (804 in LEO and 181 in MEO or GEO). NATO Member States own or operate 75% of the commercial and 3% of the military satellites in LEO, and 24% of the commercial and 16% of the military satellites in MEO or GEO.[4]

Moreover, since the advent of the digital age, cyberattacks and cyber-electronic attacks have increased roughly 10-fold, mostly conducted by State actors like Russia, China

[1] L Martino, 'Between International Politics and Technology: Dominating Cyber to Control Space?' (ISPI, 2023) <https://www.ispionline.it/en/publication/between-international-politics-and-technology-dominating-cyber-to-control-space-152123> accessed 26 March 2025.

[2] J Falcão Serra, 'Cybersecurity and Outer Space: Learning from Connected Challenges' in *Outer Space and Cyber Space: Similarities, Interrelations and Legal Perspectives* (ESPI 2021) 87.

[3] NATO, 'NATO 2022 Strategic Concept' (2022) <https://www.nato.int/nato_static_fl2014/assets/pdf/2022/6/pdf/290622-strategic-concept.pdf> accessed 26 March 2025.

[4] ESPI Launch Database.

and Iran.[5] In a 2021 study, Manulis et al. recorded 140 cyberattacks targeting critical space infrastructure that occurred between 1997 and 2020.[6] Of these, 80% targeted government, civil and military space infrastructure, while 20% attacked commercial space infrastructure.[7] The most common types of attack were theft-loss of satellite control (37%), computer network exploitation (23%), jamming (15%) and hijacking (12%). In addition to these, seizure of control (3%), eavesdropping, spoofing, and denial of service (DoS) (2%) were also recorded.[8]

Despite this rise in the frequency of attacks,[9] research has mostly addressed attacks on ground space systems[10] and only a few reported cases directly affecting a segment of space systems.

The main objective of this contribution is to identify significant gaps in the legal framework, political consensus and operational coordination, particularly when it comes to establishing thresholds for invoking collective self-defence in response to cyberattacks on space systems, and to propose a methodology to address cyberattacks against space assets within international law and NATO institutional framework.

2. EVOLUTION OF NATO'S STANCE ON THE CYBER AND SPACE DOMAINS

NATO's approach to the space and cyber domains has evolved significantly in recent decades. Although both were initially excluded from NATO's core tasks, the two domains have been gradually incorporated by NATO at different speeds, but following a similar approach. The Alliance recognized the importance of protecting its communications and command systems against cyberattacks as early as 2002, at the Prague Summit,[11] implicitly acknowledging the strategic dimension of the cyberspace domain. Instead, it took almost a decade more for the strategic relevance of space to be taken into account, and even then it was primarily as an enabler for other domains and capabilities.

[5] Space & Cyber Security, 'Space Attacks Open Database Project' <https://www.spacesecurity.info/space-attacks-open-database> accessed 26 March 2025.

[6] M Manulis, CP Bridges, R Harrison, V Sekar and A Davis, 'Cyber Security in New Space: Analysis of Threats, Key Enabling Technologies, and Challenges' (2021) 20 International Journal of Information Security 297.

[7] Space & Cyber Security, 'Space Attacks Open Database Project' <https://www.spacesecurity.info/space-attacks-open-database/> accessed 26 March 2025.

[8] The percentages presented were derived independently from the source cited in n 6 and n 7.

[9] As demonstrated by the number of studies on the matter – for example, C Poirier, *Hacking the Cosmos: Cyber Operations against the Space Sector – A Case Study from the War in Ukrain*e (Center for Security Studies ETH Zürich 2024).

[10] ibid.

[11] NATO, 'Prague Summit' (2002) <https://www.nato.int/docu/0211prague/speeches-e.pdf> accessed 26 March 2025.

As a result, the formal recognition of these domains as operational followed different timelines.

As a response to the attacks conducted against the website of NATO's Supreme Headquarters Allied Powers Europe (SHAPE) in the late 1990s amid the war in Kosovo, the attack against Estonia in 2007 and the conflict in Georgia in 2008, the 2008 Bucharest Summit Declaration emphasized the increased need for a collective defence mechanism to address cyberattacks.[12] The same year, NATO's Policy on Cyber Defence – revised in 2011 and 2014, and followed by a new Comprehensive Cyber Defence Policy in 2021[13] – formalized this shift at the operational level by focusing on protecting NATO networks, integrating cyber defence into NATO's defence planning and expanding NATO's cyber defence responsibilities to include Member States' critical infrastructure (revised version of 2014). NATO's 2010 Strategic Concept included cyber threats as a direct challenge to transatlantic and national security. Again, similar recognition for space came nearly a decade later, with the 2018 Brussels Summit Declaration explicitly acknowledging the importance of space for the security of the Alliance and its operational effectiveness, with space being recognized as the fifth operational domain of the Alliance one year later.[14]

Three years later, the 2022 Overarching Space Policy cemented this shift and provided a policy framework for NATO's approach to space security, including ensuring resilience and cooperation among Member States.

In both cases, the approach was conservative. The fact that space was declared an operational and not a warfighting domain is not merely a matter of formality. Unlike some Member States of the Alliance that have expressly qualified space as a warfighting domain (e.g. the US in 2020), by declaring it an operational domain and conceiving of it as an enabler of operations, NATO did not intend to focus on the ability to deny its adversaries access to space or on developing and deploying NATO-owned space capabilities. Rather, the identification of space as an operational domain explicitly recognized the role that outer space plays in military operations and for national security, while aiming to achieve greater integration and interoperability among space infrastructures belonging to different Member States.[15]

The same considerations apply to the recognition of both domains as areas of possible activation of the collective defence clause. For cyber, a turning point was the 2014 Wales Summit, where NATO declared cyber defence a part of its core collective defence

[12] NATO, 'Bucharest Summit Declaration' (2008) <https://www.nato.int/cps/en/natohq/official_texts_8443.htm> accessed 26 March 2025.

[13] NATO, 'Cyber Defence' (2024) <https://www.act.nato.int/activities/cyber/> accessed 26 March 2025.

[14] NATO, 'London Declaration' (2019) <https://www.nato.int/cps/en/natohq/official_texts_171584.htm> accessed 26 March 2025.

[15] A Stickings, 'Space as an Operational Domain: What Next for NATO?' (October 2020) 40(91) *RUSI Newsbrief* <https://static.rusi.org/stickings_web_0.pdf> accessed 26 March 2025.

task under Article 5 of the North Atlantic Treaty (Washington Treaty)[16] in case of a lethal large-scale cyberattack targeting one of its members, commitment reinforced two years later with the Cyber Defence Pledge, issued at the Warsaw Summit of 2019. This facilitated cooperation among NATO Member States to improve national cyber resilience.[17] Space received similar recognition at the 2021 Brussels Summit, but the approach was more cautious and less integrated into the NATO framework than the cyber domain was, perhaps due to the Alliance's lack of direct ownership of space capabilities, as opposed to the proprietary information and computer networks used in its military missions, as well as the dual-use nature of many space systems.[18]

Finally, even in terms of the development of institutional operational frameworks, NATO's cyber defence posture was institutionalized long before its stance on space. First with a dedicated Cyber Defence Management Authority (CDMA) in 2008,[19] followed by the creation in 2012 of a specific agency dedicated to cyber defence at SHAPE Headquarters. The NATO Communications and Information Agency (NCIA) began hosting the NATO Cyber Security Operations Centre (CyOC) in 2016, while a research and education centre was created in 2008,[20] the NATO Cooperative Cyber Defence Centre of Excellence, to provide expertise and conduct exercises involving NATO allies and partners.[21] In addition, plans have been formulated to establish an Integrated Cyber Defence Centre by 2028.[22] Dedicated structures for space were established, starting with the NATO Space Centre at NATO's Allied Air Command in 2020,[23] followed by the NATO Space Centre of Excellence in 2023.[24] However, these primarily focus on situational awareness, resilience and the protection of space-based assets rather than on offensive counterspace operations.

This section shows how cybersecurity has become deeply embedded within NATO's core activities, institutional structures and mechanisms, and space has also been integrated using a similar methodology and at a faster pace, albeit superficially for now.

[16] NATO, 'Wales Summit Declaration' (2014) <https://www.nato.int/cps/en/natohq/official_texts_112964.htm> accessed 26 March 2025.

[17] NATO, 'Cyber Defence Pledge' (2016) <https://www.nato.int/cps/en/natohq/official_texts_133177.htm> accessed 26 March 2025.

[18] NATO, 'Brussels Summit Communiqué' (2021) <https://www.nato.int/cps/en/natohq/news_185000.htm?selectedLocale=en> accessed 26 March 2025.

[19] Fondation pour la recherche stratégique, 'Is NATO Ready for Cyber War?' (2021) <https://frstrategie.org/en/publications/nato-briefs-series/nato-ready-cyber-war-2021> accessed 26 March 2025.

[20] CCDCOE, 'About Us' <https://ccdcoe.org/about-us/> accessed 26 March 2025.

[21] Exercises are also developed under the Cyber Coalition, while NATO has established a structured dialogue with industry and computer firms through the NATO Industry Cyber Partnership (NICP)..

[22] Breaking Defence, 'NATO to Launch New Cyber Center by 2028: Official' (2024) <https://breakingdefense.com/2024/12/nato-to-launch-new-cyber-center-by-2028-official/> accessed 26 March 2025.

[23] NATO Allied Air Command, 'NATO Agrees New Space Centre at Allied Air Command' (2020) <https://ac.nato.int/archive/2020/NATO_Space_Centre_at_AIRCOM> accessed 26 March 2025.

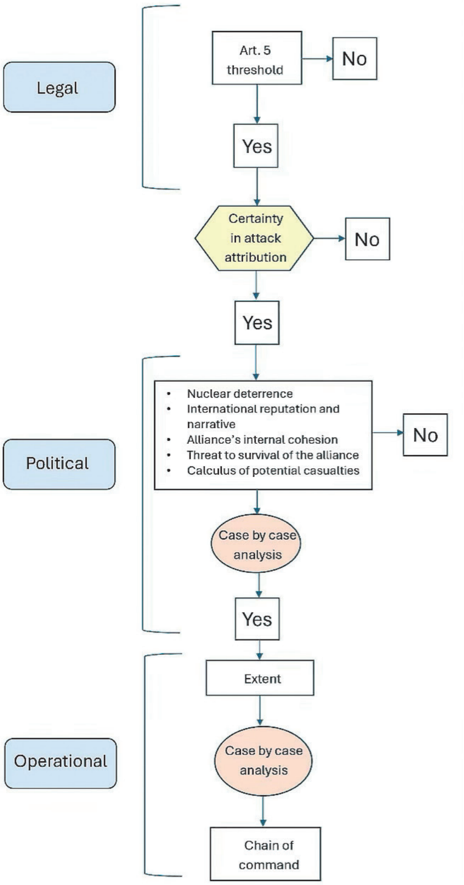
[24] NATO Space CoE, 'Signing of the Memorandum of Understanding Establishing the NATO Space COE' (2023) <https://space-coe.org/signing-of-the-memorandum-of-understanding-establishing-the-nato-space-centre-of-excellence/> accessed 26 March 2025.

However, it also reveals a notable absence of integration, coordination or communication channels between the two domains at the structural, policy, exercise, planning and doctrinal levels.

3. THE DIFFERENT LAYERS INVOLVED IN THE EVALUATION

The complexity of the relationship between the cyber and space domains is also revealed in the determination of the reaction by States and within NATO to cyber threats against the space systems of Member States.
The authors suggest breaking down the evaluation into three layers – the legal, political and operational (see Figure 1).

FIGURE 1: METHODOLOGY FOR A CASE-BY-CASE EVALUATION


Legal
Art. 5 threshold
No
Yes
Certainty in attack attribution
No
Yes
Political
Nuclear deterrence
International reputation and narrative
Alliance's internal cohesion
Threat to survival of the alliance
Calculus of potential casualties
No
Case by case analysis
Yes
Operational
Extent
Case by case analysis
Chain of command

Given the expected increase in hostile cyber events targeting space infrastructure, a structured methodology guiding identification on a case-by-case basis would, over time, help streamline assessments at the legal, policy and operational levels, and would establish a clearer threshold for triggering Article 5.

A. The Legal Layer and the Applicable Regime

The first step of the assessment relates to the legal qualification of the event under examination and, consequently, a potential activation of Article 5 of the Washington Treaty. For this, international space law and international law should be considered.

The outer space environment represents a unique legal domain, governed by a *sui generis* framework that prioritizes self-restraint in weapons development and deployment.[25] This regulatory approach is characterized by a specialized legal framework designed to address on a norm-to-norm basis the peculiarities of space.

In situations where this framework fails to adequately address issues such as the use of force in outer space, attention must shift to broader principles of international law, which serve as the *lex generalis* in relation to space law. This interpretation is supported by Article III of the Outer Space Treaty, underscoring the foundational role of international law in regulating outer space activities and aligning space law with the overarching legal framework established by the UN Charter. The priority of the UN Charter in case of legal conflict is further cemented by Article 103 of the UN Charter[26] and Article 30 of the Vienna Convention, which ensure the primacy of the UN Charter in governing international relations, including in outer space.[27]

Once it is established that the legal regime governing the use of force under the UN Charter applies to outer space, examining how the use of force is conceptualized and regulated in this unique domain becomes critical. The cornerstone principle on the prohibition of the use of force is articulated in Article 2(4) of the UN Charter, recognized as a norm of customary international law and attaining the status of *jus cogens*, i.e. meaning it is a peremptory norm from which no derogation is permitted,[28] as reaffirmed by the International Court of Justice (ICJ) in its 1986 judgment in *Nicaragua v USA*.[29] This means that the prohibition against the threat or use of force is universally binding

[25] Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, including the Moon and Other Celestial Bodies (Outer Space Treaty), art IV.

[26] Charter of the United Nations, art 103: 'In the event of a conflict between the obligations of the Members of the United Nations under the present Charter and their obligations under any other international agreement, their obligations under the present Charter shall prevail.'

[27] Vienna Convention on the Law of Treaties, art 30: '1. Subject to article 103 of the Charter of the United Nations, the rights and obligations of States parties to successive treaties relating to the same subject-matter shall be determined in accordance with the following paragraphs. 2. When a treaty specifies that it is subject to, or that it is not to be considered as incompatible with, an earlier or later treaty, the provisions of that other treaty prevail.'

[28] JA Frowein, 'Jus Cogens' in R Bernhardt (ed), Encyclopedia of Public International Law (Springer 1997) vol III, 65.

[29] *Case Concerning Military and Paramilitary Activities in and against Nicaragua (Merits)* ICJ <https://www.icj-cij.org/files/case-related/70/070-19860627-JUD-01-00-EN.pdf> accessed 26 March 2025.

and applies to all States individually and as part of international setups, regardless of whether they are party to the UN Charter. However, questions have arisen regarding the territorial scope of this prohibition, particularly as regards outer space, which is a non-territorial domain, although arguments favouring a narrow interpretation of the term ‘territorial integrity’ in Article 2(4) have been mostly rejected.[30] Instead, the reference to territorial integrity must be understood in conjunction with the broader mandate of Article 2(4), which prohibits force used ‘in any other manner inconsistent with the Purposes of the United Nations’. The purposes of the UN include maintaining international peace and security, and promoting peaceful relations among States. Therefore, any interpretation that limits the application of Article 2(4) to terrestrial domains undermines these fundamental goals, an interpretation reinforced by the ICJ in the Corfu Channel case,[31] which emphasized the need to interpret international law considering its purpose and principles.

Applying this reasoning to outer space, it can be argued that the prohibition on the use of force extends to activities involving space objects.[32] In this regard, the principle of non-use of force applies to actions targeting space objects or activities that would otherwise violate the peace and security objectives of the UN, including cyberattacks (consistent with the approach taken by the authors of the Tallinn[33] and the Woomera manuals[34]).

However, the interpretation of the prohibition on the use of force must align with the unique characteristics of outer space. For example, the destruction or disabling of a satellite could have far-reaching implications for global stability, as satellites play critical roles in communication, navigation and national security. Moreover, unilateral use of force against space objects would not only disrupt the peaceful use of outer space but could also escalate conflicts on Earth, contradicting the purposes of the UN.

B. Possible Elements for Evaluation of the Threshold of an Armed Attack in Space and NATO’s Contribution

As demonstrated, the prohibition on the use of force under international law extends to outer space. This includes not only the foundational principles of *jus ad bellum* but also its two major exceptions: self-defence and Security Council authorization. Through a teleological interpretation of the relevant provisions of international law, the extension of self-defence to outer space is justifiable, provided such actions comply with the limitations outlined in Article IV of the Outer Space Treaty. The UN Charter seeks to maintain international peace and security, and Article 51 provides

[30] See eg DW Bowett, *Self-Defence in International Law* (Manchester University Press 1958).

[31] *The Corfu Channel Case (Merits)* ICJ <https://www.icj-cij.org/files/case-related/1/001-19490409-JUD-01-00-EN.pdf> accessed 26 March 2025.

[32] F Tronchetti, ‘The Right of Self-Defence in Outer Space: An Appraisal’ (2014) 63 ZLW 92, 98.

[33] MN Schmitt, *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (CUP 2017).

[34] J Beard and D Stephens (eds), *The Woomera Manual on the International Law of Military Space Operations* (Oxford Academic 2024).

a legal mechanism for States to respond to armed attacks, including those taking place in space or involving space systems. In line with this interpretation, States have increasingly begun to refer to the possibility of resorting to self-defence in space, as evidenced by the number of national space doctrines that have proliferated over the past two decades.

Article 51 establishes a clear condition for invoking self-defence, namely the occurrence of an armed attack, but does not provide a precise definition of armed attack. This lack also reverberates on the meaning of armed attack for the purposes of activating Article 5 of the Washington Treaty. Without delving into the extensive literature on the *criteria ratione personae, materiae*, or *temporis* that determine the existence of an armed attack,[35] this paper will build on the ICJ's specification that not all uses of force qualify as armed attacks, entailing a distinction between 'grave' and 'less grave' uses of force, where only the former – those that exceed a significant threshold of scale and effect – constitute armed attacks.[36] Along these lines, an armed attack generally involves 'the use of arms or military force of offensive, destructive, and illegal nature',[37] with such attacks being of sufficient magnitude to compel the victim State to respond in self-defence, as inferred from the regime of proportionality in self-defence.[38]

In the context of outer space, this definition must be applied to both irreversible and reversible operations, including those conducted through cyber means.

In case of irreversible operations, such as the use of kinetic and physical counterspace operations to destroy space objects, the assessment of the existence of an armed attack could be slightly clearer due to their destructive and offensive nature.[39]

With reversible operations, however, attribution and assessment of the scale and effects in interdependent or dependent infrastructures is more complex, making it difficult to decide whether the offensive action is grave enough to trigger a response. Especially in the case of cyber operations that only temporarily disable a space infrastructure without destroying it, the determination of a threshold for whether this amounts to an armed attack becomes particularly challenging, as does the decision on how Member States should respond. Here, one criterion to evaluate the impact of the attack could

[35] For an early work on the subject, see T Ruys, *'Armed Attack' and Article 51 of the UN Charter: Evolutions in Customary Law and Practice* (CUP 2010).

[36] *Case Concerning Military and Paramilitary Activities in and against Nicaragua (Merits)* (n 33) para 195; *Case Concerning Oil Platforms (Islamic Republic of Iran v United States of America)*, para 51.

[37] F Tronchetti, 'The Right of Self-Defence in Outer Space: An Appraisal' (2014) 63 ZLW 92, 98.

[38] A Cassese, *International Law* (OUP 2005) 355.

[39] In line with recent developments in space law, these actions also violate space law, as they jeopardize the freedom of other States to use outer space peacefully, as established in arts I, III and IX of the Outer Space Treaty.

be the distinction between critical and non-critical space infrastructure at the national level, as well as whether the target is critical to NATO's operations as an alliance.[40]

Critical space infrastructure (CSI) is critical to interconnected systems on Earth. Counterspace operations disrupting or incapacitating these infrastructures can trigger cascading failures of multiple dependent systems, such as electrical or telecommunication networks, regardless of whether the disruption is temporary or permanent. This influences the assessment of scale and effects in determining a response threshold.

However, different NATO Member States have different dependencies and priorities with respect to space infrastructure, potentially leading to misaligned assessments of cyber threats to these systems. Therefore, at the NATO level, a first step could be to promote a classified internal dialogue on national approaches and criteria for designating specific space infrastructures as critical, recognizing their vulnerability to disruption.[41] In addition, to evaluate the scope and effects of an attack, exercises simulating complex attack scenarios involving multi-domain operations could help assess the nature of the targeted infrastructure and its externalities based on different scenarios. They could also help to evaluate the temporal element of the attack within broader incapacitation operations, the fortuity or intentionality of the attack (including considering its repetitiveness and context),[42] as well as the circumstances surrounding it in space (through Space Domain Awareness) and on Earth.

NATO could therefore serve as a forum to build consensus on the types of space infrastructure considered critical, regardless of individual States' actual or future dependence on them, supporting more consistent assessments on a case-by-case basis as well as facilitating the creation of protection and resilience tools for these infrastructures over the long term.[43] In addition, it would also facilitate the development and implementation of best practices by establishing minimum security standards and simplifying risk and threat assessments across the Alliance. Finally, such exercises and associated identified protocols should also consider different types of actors involved in operations, including commercial actors supporting operations or providing services to NATO Member States.

[40] G Pavesi, 'Legal Management of the Concept of Risk in Reversible Operations Against Space Assets' in *Legal Developments in Cybersecurity and Related Fields* (Springer 2024).

[41] G Pavesi, 'NATO versus Non-kinetic Threats: Implications and Opportunities' (Centre for International Governance Innovation, 29 January 2023) <https://www.cigionline.org/articles/nato-versus-non-kinetic-threats-implications-and-opportunities/> accessed 27 March 2025.

[42] F Tronchetti, 'The Right of Self-Defence in Outer Space: An Appraisal' (2014) 63 ZLW 92, 117.

[43] See eg B Unal, *Cybersecurity of NATO's Space-Based Strategic Assets* (Chatham House 2020).

4. THE POLITICAL LAYER OF THE EVALUATION

Once the legal assessment is concluded, the political variables should be addressed, particularly the ones impacting the Alliance's internal cohesion and wider global security externalities.

On the one hand, NATO's approach to determine the activation of Article 5 on a case-by-case basis is itself a foundational element of deterrence, leaving the attacker in a situation of uncertainty with respect to the expected response to a cyberattack against a space infrastructure. This allows the Alliance to keep its options open and determine, based on the circumstances, whether to intervene. Operationally and politically, this ambiguity and lack of predefined thresholds therefore reinforce NATO's deterrence, avoiding creating vulnerabilities by setting specific thresholds.

However, in the long run, this ambiguity could also result in a lack of clarity in NATO's stance towards these threats and a lack of consistency of action, especially in the face of the expected increase in the frequency of such situations. Therefore, should NATO decide to act in a specific case, one element that should be addressed is the political narrative that accompanies the response to the threat, with the goal being to maintain the Alliance's legitimacy on the global stage. As mentioned earlier, a cyberattack on space infrastructure brings up several legal considerations, particularly concerning the attribution of the attack and the proportionality of the response.[44] Regarding the latter, should it decide to act, politically it might be advisable for NATO to transparently communicate evidence of the attack, and at the same time emphasize the Alliance's commitment to proportionality and the rule of law, to ensure its international credibility.[45] To achieve greater multilateral consensus, an agreement of intent with international organizations or non-NATO States might be necessary so as to present the response as part of a broader effort to support international stability and security.[46]

Instead, a weak or non-cohesive response to a cyberattack against space infrastructures could undermine the credibility of the Alliance, encouraging new attacks and questioning the relevance of Article 5 in modern conflicts.[47] This is crucial to consider, as nations with advanced space and cyber capabilities may push for a less proportionate response, while others may invoke conservatism to avoid escalation,

[44] JA Lewis, *Creating Accountability for Global Cyber Norms* (Center for Strategic and International Studies 2022).

[45] DP Fidler, R Pregent and A Vandurme, 'NATO, Cyber Defence, and International Law' (2013) 4(1) St. John's Journal of International & Comparative Law.

[46] NATO, 'Relations with the United Nations' (2023) <https://www.nato.int/cps/en/natohq/topics_50321.htm> accessed 27 March 2025.

[47] ED Lonergan and SB Moller, 'NATO's Credibility Is on the Line with Its Cyber Defence Pledge. That's a Bad Idea.' (2022) <https://www.politico.com/news/magazine/2022/04/27/nato-credibility-cyber-defense-pledge-russia-ukraine-00027829> accessed 27 March 2025.

thus causing an impasse.[48] This problem can be overcome by establishing clearer internal perimeters of the criteria that determine the exceeding of thresholds, so as to avoid disagreements and guarantee unified action. In this sense, a decisive but proportionate response would reaffirm NATO's commitment to collective defence, while signalling that the Alliance remains capable of addressing today's threats.[49]

At this stage of reflection, however, a critical question emerges: Do States today possess the political will to address these issues at the NATO level? Table I, focused on the cyber domain, shows that in just two years, States adopted different approaches in responding to three different events. In the case of the Colonial Pipeline cyberattack, the US reaction[50] was to signal to adversaries a list of *off-limits targets*, indirectly contributing to the definition of a future threshold.[51] The second case, that of the Viasat hack, generated much academic debate,[52] but the affected State nonetheless declared that 'the satellite outage was a really huge loss in communications *in the very beginning of war'*.[53] The US said that 'Russia launched cyberattacks in late February against commercial satellite communications networks to disrupt Ukrainian command and control *during the invasion*, and those actions had spillover impacts into other European countries'.[54] This statement makes the timing of the attack a determining factor in the legal evaluation of the event. Finally, in the case of the cyberattack on Albania, despite the victim State's push to invoke Article 5, *political considerations* trumped the activation of the clause.[55]

[48] NATO, 'Consensus Decision-Making at NATO' (2023) <https://www.nato.int/cps/en/natohq/topics_49178.htm> accessed 27 March 2025.

[49] NATO, 'Warsaw Summit Communiqué' (2016) <https://www.nato.int/cps/en/natohq/official_texts_133169.htm> accessed 27 March 2025.

[50] The White House, 'Remarks by President Biden on the Colonial Pipeline Incident' (2021) <https://www.whitehouse.gov/briefing-room/speeches-remarks/2021/05/13/remarks-by-president-biden-on-the-colonial-pipeline-incident/> accessed 27 March 2025.

[51] V Soldatkin and H Pamuk, 'Biden Tells Putin Certain Cyberattacks Should Be "Off-Limits"' (*Reuters*, 17 June 2021) <https://www.reuters.com/technology/biden-tells-putin-certain-cyber-attacks-should-be-off-limits-2021-06-16/> accessed 28 March 2025.

[52] Ukraine Symposium, 'The Risk of Commercial Actors in Outer Space Drawing States into Armed Conflict' (Lieber Institute West Point) < https://lieber.westpoint.edu/commercial-actors-outer-space-armed-conflict/> accessed 28 March 2025.

[53] D Cattler and D Black, 'The Myth of the Missing Cyberwar' (*Foreign Affairs*, 6 April 2022) <https://www.foreignaffairs.com/articles/ukraine/2022-04-06/myth-missing-cyberwar> accessed 28 March 2025.

[54] US Department of State, 'Attribution of Russia's Malicious Cyber Activity Against Ukraine' (2022) <https://www.state.gov/attribution-of-russias-malicious-cyber-activity-against-ukraine/> accessed 28 March 2025.

[55] M Miller, 'Albania Weighed Invoking NATO's Article 5 Over Iranian Cyberattack' (*Politico*, 10 May 2022) <https://www.politico.com/news/2022/10/05/why-albania-chose-not-to-pull-the-nato-trigger-after-cyberattack-00060347> accessed 28 March 2025.

TABLE I: ‘TRADITIONAL’ CYBERATTACKS

Attack	Year	Description	Legal and political reactions
Colonial Pipeline ransomware attack	2021	Russian cybercriminal group DarkSide launched a ransomware attack on Colonial Pipeline, the largest US pipeline operator, disrupting operations. This led to fuel shortages and widespread inefficiencies along the East Coast.	US President Biden handed Russian President Vladimir Putin a list of 16 US critical infrastructure sectors that are off-limits to any Russian cyberattack. Despite this, the US administration did not turn to NATO and did not openly discuss Article 5. Rather, Washington chose to tackle the issue bilaterally.
Viasat hack	2022	Russian military intelligence (GRU) targeted Viasat, disabling around 20,000 modems and disrupting internet access across Ukraine and Europe. The attack also crippled communications for Ukraine’s military, police and intelligence services.	Despite the spillover effects of the attack and the indiscriminate targeting of Viasat modems in the context of an international armed conflict, the Alliance did not publicly deliberate the application of Article 5.
Cyber campaign against Albania	2022	Albania suffered a major cyberattack attributed to four alleged Iranian government APT actors. The attackers disrupted key government service filtrated and leaked sensitive data and temporarily disabled border control systems.	The Albanian government severed all diplomatic relations with Tehran – the first time after a cyberattack. Then the Albanian government also discussed turning to NATO’s Article 5. However, Prime Minister Edi Rama eventually decided against turning to NATO, noting that he has too much respect for his friends and Allies to tell them what to do.

The same observations are valid in relation to the space domain. Table II shows that there was consistency of action – or rather inaction – in this case as well, but no discussions were held on activating the collective defence clause, potentially contributing to setting a ‘higher’ threshold in relation to cyber operations targeting space systems at the national and NATO level.

TABLE II: CYBERATTACKS AGAINST SPACE SYSTEMS

Attack	Year	Description	Legal and Political Reactions
Russia suspected of jamming GPS in Finland and Norway	2018	During NATO's Trident Juncture exercises in Scandinavia, Finland and Norway reported GPS signal disruptions, posing air safety risks.	Finnish Prime Minister Juha Sipilä suggested the jamming was deliberate and Russia was likely to be behind it, given its known electronic warfare capabilities. Anyway, no discussion was held due to the fact that at the time Finland was not in NATO.
Russia's GRU attack on Viasat	2022	Russian military intelligence (GRU) targeted Viasat, disabling around 20,000 modems and disrupting internet access across Ukraine and Europe. The attack also crippled communications for Ukraine's military, police, and intelligence services.	Despite the spillover effects of the attack and the indiscriminate targeting of Viasat modems in the context of an international armed conflict, the Alliance did not publicly deliberate the application of Article 5.
Alleged Russian spoofing against Finland in the Baltic Sea	2024	Finland has faced ongoing satellite navigation disruptions involving the spoofing of global navigation satellite systems, which are critical to maritime navigation. These interferences have caused vessels to lose their bearings, increasing accident risk.	Authorities suspected Russia was behind the disruptions, but no discussions were held and no countermeasures were taken.

These examples only corroborate the fact that, given there are no clear thresholds for cyberattacks on Earth, it is, to say the least, complicated to deal with such cases in space, for which the international legal framework is at present rather rudimentary.

Inconsistencies in the reactions to cyberattacks across the two domains can also be seen in practice. An analysis of significant events reveals that different priorities at the time of the evaluation led to different reactions by Member States and shaped their subsequent courses of action.

With this in mind, it should be of paramount importance to stimulate critical discussion within NATO on how the Alliance intends to define its role in a changing security landscape. Indeed, without a clearer frame of reference, NATO risks further ambiguity that could weaken its credibility and deterrence capacity.

On the other hand, these challenges represent an opportunity. NATO can use the current strategic environment to revitalize itself by adapting to contemporary threats in the cyber, space and hybrid warfare domains. By proactively shaping a coherent and unified posture, NATO can strengthen its relevance and cohesion, ensuring that it remains able to cope with the complex security dynamics of the 21st century.

5. THE OPERATIONAL LAYER OF THE EVALUATION

In an interview with *SpaceNews* released immediately after the Russian attack on Viasat, the commander of the US Space Force's Space Operations Command, Lt. Gen. Stephen Whiting, admitted that 'cyberspace is the soft underbelly of our global space networks'.[56] This statement can also be applied to NATO, whose decentralized command structures and multinational composition make it difficult to create unified approaches to interdimensional operations. Therefore, ensuring that information flows and that decision-making and coordination occur quickly and effectively across domains remains a continuing concern. Since its institution, NATO has been organized and structured around domain-specific operation centres. However, these 'traditional' domains differ substantially from the space and cyber domains, which are less geographically tied, rely more on commercial infrastructure and often include actors beyond States' armed forces. These peculiar physical conditions and the increasing interconnectedness between the two domains require an integrated, cross-domain operational approach.[57]

Despite the establishment of space and cyber operational entities, the fragmentation of command structures still does not allow NATO to properly counter threats within these operational domains. This gap is caused by the lack of a centralized command framework and the operational divergence among Member States, which weakens overall interoperability across domains and impedes the implementation of a cohesive strategy. For instance, NATO in the past has sometimes found it difficult to coordinate responses within its domains – first in the 'traditional' ones and then in the 'emerging' ones – due to cultural, doctrinal and technological disparities.

Moreover, the absence of uniform standards for multi-domain operations, coupled with technological gaps and differing economic capacities of Member States, further complicates the issue.[58]

These operational problems are also reflected in the purely technical aspects of the cybersecurity of NATO's CSI. The use of outdated systems by less technologically advanced members of the Alliance and insufficiently integrated regulatory frameworks further weaken its position.

[56] S Erwin, 'Space Force to Shore Up Cybersecurity as Threats Proliferate' (*Space News*, 6 April 2022) <https://spacenews.com/space-force-to-shore-up-cybersecurity-as-threats-proliferate/> accessed 28 March 2025.

[57] L Caprio, M Garcia Flores, RA Grassi and C Toti, *NATO Multi-Domain Operations: Challenges for the European Land Forces* (FINABEL 2024).

[58] ibid.

Finally, misalignment in data classification levels between NATO and its Member States and the transmission of unencrypted data across systems (mostly due to the different technological levels of Member States) amplify the risks in the supply chain.[59]

Summing up, while a fundamental change in NATO's approach is required to overcome these challenges, the upside is that, at both the political and operational levels, the Alliance already has the potential to effectively address the threats discussed in the article, as it already possesses mechanisms for interoperability[60] and the secure transmission of classified information. The Alliance's existing frameworks for intelligence sharing and joint operations,[61] although primarily developed for traditional domains, can be adapted to space and cyber operations.

At the operational level, the space and cyber domains are still being addressed in silos, without effective mechanisms for dialogue and coordination. Using scenarios based on actual events, such as those briefly discussed in Tables I and II, NATO could explore how to integrate institutional and operational arrangements for each domain. This would include establishing clear chains of command across the Alliance as well as efficient communication channels in case of threats or attacks.

6. CONCLUSIONS: A METHODOLOGY FOR FUTURE ACTION

This paper has examined the evolving interplay between cyber and space security within NATO, identifying key gaps in legal frameworks, policy coordination and operational structures. The analysis has demonstrated that while cyber threats have been integrated into NATO's strategic and operational planning over the past two decades, space – despite being recognized as an operational domain – has yet to receive the same level of institutional integration. The discussion also highlighted the fragmented nature of legal and political responses to cyberattacks on space assets, illustrating that while NATO acknowledges the risks, there remains a lack of clear thresholds and, for now, political will to invoke collective defence mechanisms under Article 5 of the Washington Treaty.

At the legal level, the assessment outlined the applicability of international law to space-based cyber threats, emphasizing the need to reconcile space law's norm-to-norm approach with the broader principles of international law, particularly regarding

[59] BK Vollmer, *NATO's Mission-Critical Space Capabilities Under Threat: Cybersecurity Gaps in the Military Space Asset Supply Chain* (Paris School of International Affairs 2021).

[60] NATO, 'Interoperability: Connecting Forces' (2023) <https://www.nato.int/cps/en/natohq/topics_84112.htm> accessed 28 March 2025.

[61] NATO, 'Joint Intelligence, Surveillance and Reconnaissance' (2024) <https://www.nato.int/cps/en/natohq/topics_84112.htm> accessed 28 March 2025.

jus ad bellum. The study demonstrated that while kinetic attacks on space assets would likely meet the threshold for an armed attack, cyber operations present a more complex challenge due to their reversible nature and the interdependent nature of CSI. The absence of a clear, NATO-wide classification of critical space assets further complicates the assessment of proportionality and response mechanisms.

The political evaluation underscored that while NATO has progressively integrated cyber defence into its core tasks, the decision to invoke collective defence remains contingent upon political consensus among its Member States. Past responses to cyberattacks, both on the terrestrial and space targets, have been inconsistent, with political considerations often outweighing legal and strategic imperatives. The lack of a unified threshold for cyberattacks on space systems has contributed to uncertainty regarding NATO's stance, potentially encouraging adversaries to exploit these ambiguities.

At the operational level, this contribution highlighted structural limitations within NATO's command and control framework. Despite the establishment of dedicated cyber and space entities, the lack of an integrated, cross-domain approach continues to hinder NATO's ability to respond effectively to cyber threats targeting space assets. Technological disparities among Member States and outdated cybersecurity measures further exacerbate these vulnerabilities, leaving NATO's CSI exposed to emerging threats.

At all levels analysed, the Alliance is demonstrating progress and doing so at a decidedly fast pace in reaction to the changing operational environment. However, in facing the modern challenges that arise from the intersection of the two domains of cyber and space, the Alliance needs to develop a structured methodology internally to understand and define the perimeter of a case-by-case evaluation of the activation of Article 5, through an integrated and holistic approach that considers multiple levels of determination of each variable involved at the various levels of the decision-making process.

CyCon 2025: The Next Step
17th International Conference on Cyber Conflict
C. Kwan, N. Gratzer, K. Podiņš, M. Tolppa (Eds.)
2025 © NATO CCDCOE Publications, Tallinn

Permission to make digital or hard copies of this publication for internal use within NATO and for personal or educational use when for non-profit or non-commercial purposes is granted provided that copies bear this notice and a full citation on the first page. Any other reproduction or transmission requires prior written permission by NATO CCDCOE.

Police Your Own: Establishing an "Unwilling or Unable" Analog for Deploying Collective Countermeasures in Cyberspace

Gwendolyn Lee Strasberg
Marine Forces Cyberspace Command
United States

Andom Gherezghiher
Marine Forces Cyberspace Command
United States

Abstract: In today's digital age, each nation's capacity to police its own cyber infrastructure is crucial to maintaining stability across the global ecosystem. Interconnected networks subvert traditional notions of territorial sovereignty. Apathy toward (or support of) malicious actors by any single State threatens the broader international community in cyberspace. Accordingly, vast disparities in cyber capabilities and responsiveness render collective action a valuable resource for digitally vulnerable States. Limiting States to collective self-defense for third-party intervention pushes them to aggressively and belligerently characterize incidents to elicit third-party assistance. Establishing an alternate mechanism for joint action against nefarious activity via lower-threshold collective countermeasures would result in a more secure and organized global network. Despite major endorsements, however, there remains a lack of international consensus regarding the permissiveness or prohibition of collective countermeasures. Given the vast potential for policing cyberspace, establishing criteria for the use of collective countermeasures is critically necessary in the development of cyberlaw. In pursuit of that development, this work traces collective self-defense and collective countermeasures to a common legal foundation in the prohibition of acts of aggression and resulting due diligence obligations. After identifying support in international law, this paper introduces a unifying four-factor test for deploying collective countermeasures, thereby operationalizing the concept and establishing a clear, replicable threshold for implementation in cyberspace.*

Keywords: *collective countermeasures, cyberspace, due diligence, erga omnes obligations, United Nations Charter*

* The views and opinions presented herein are those of the authors and do not necessarily represent the views of DoD or the United States Marine Corps. Discussions of, or reference to, any interpretations of the law do not constitute DoD endorsement of those positions. The appearance of external links does not constitute DoD endorsement of the linked websites, or the information therein.

1. INTRODUCTION

By and large, the front line of modern conflict has shifted from the physical/kinetic arena to non-kinetic actions upon a network of interconnected systems, perpetually at risk of unrelenting cyber assault. Yet, the swift emergence of the technological battlefield is in sharp contrast to the scholarly inertia in the legal obligations directing and regulating it, rendering State practice and *opinio juris* chronically disconnected from modern operations. This anachronistic model of law creates gaps and irrational incentive structures, hindering the adaptation of international law to cyberspace. In a world where cyber threats transcend borders, the failure of any one State to secure its digital infrastructure and society can destabilize the entire global order.

In this context, countermeasures are a compelling tool for policing cyber terrain below the use-of-force threshold for self-defense.[1] Countermeasures allow an "injured" State, which has suffered a violation of an international obligation, to take otherwise unlawful actions against the State responsible for causing the harm.[2] The Articles on State Responsibility set forth the conditions required for compliance with international law:

1. An injured State 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 part two.
2. Countermeasures are limited to the non-performance for the time being of international obligations of the State taking the measures toward the responsible State.
3. Countermeasures shall, as far as possible, be taken in such a way as to permit the resumption of performance of the obligations in question.[3]

In other words, countermeasures must be reversible and cease promptly once the responsible State resumes compliance.[4] The scale of the countermeasure must not be excessive when compared to the harm suffered.[5]

Countermeasures are an important and established feature of international law, described as a "centerpiece… of self-help" in modern international relations.[6] The *Tallinn Manual 2.0* says that in cyberspace, a "State may be entitled to take

[1] Michael Schmitt, *Lieber Institute White Paper: Responding to Malicious or Hostile Actions under International Law*, ARTICLES OF WAR (Apr. 26, 2022), https://lieber.westpoint.edu/white-paper-responding-malicious-hostile-actions-international-law/.

[2] G.A. Res. 56/83, Articles on Responsibility of States for Internationally Wrongful Acts (Jan. 28, 2002) [hereinafter *Articles on State Responsibility*].

[3] *Id*. at 13 ("Article 49. Object and limits of countermeasures").

[4] *Id*. at 14 ("Article 53. Termination of Countermeasures").

[5] *Id*. at 13 ("Article 50. Proportionality").

[6] Michael N. Schmitt & Sean Watts, *Collective Cyber Countermeasures*, 12 HARV. NAT'L SEC. J. 373, 380–85 (2021).

countermeasures, whether cyber in nature or not, in response to a breach of an international legal obligation that it is owed by another State." Yet not every State is equally equipped to address threat actors in cyberspace.

Technological asymmetry renders collective countermeasures—where a third-party State takes countermeasures on behalf of an injured State—an appealing development in the doctrine.[7] Although collective action has been accepted in self-defense, stakeholders remain divided on endorsing a collective approach to countermeasures. While technologically advanced States like France and Canada object to their legality, other actors (especially past victims of cyberattacks) and prominent academics view them as necessary to protect vulnerable States. Notwithstanding, if current international law were to exclusively favor collective action in self-defense, it would create a perverse incentive to overclassify incidents for the purpose of engaging third-party assistance. Given their capacity to cure this misalignment and meet practical demands in cyberspace, momentum favors the acceptance of collective countermeasures by international law.

This progressive development could result in erratic, undesirable applications of the doctrine unless unified criteria are in place for evaluating collective countermeasures. Indeed, creating a replicable mechanism for collective action below the use-of-force threshold could prevent unnecessary and escalatory characterization and address the asymmetry in cyber capacity that makes small States appealing targets for malicious actors in cyberspace.

By drawing upon the successful unwilling or unable model, collective self-defense provides a useful foundation for collective countermeasures. The roots that both doctrines share in the prohibition of acts of aggression make collective self-defense a helpful platform from which to operationalize a clear, replicable threshold for implementing collective countermeasures.

This paper seeks ultimately to develop the law, address the risk posed by asymmetrical cyber capabilities, and respond to the demands of modern society. After building on past scholarship contemplating the legal justifications for collective countermeasures in cyberspace, it advocates a four-factor test derived from the unwilling or unable test in collective self-defense. The test considers the following: (1) consent and cooperation

[7] *See, e.g.*, Przemyslaw Roguski, *Collective Countermeasures in Cyberspace - Lex Lata, Progressive Development or a Bad Idea?*, in 12TH INTERNATIONAL CONFERENCE ON CYBER CONFLICT (CYCON) 1300, NATO CCDCOE 25 (2020), https://doi.org/10.23919/CyCon49761.2020.9131715; Oona Hathaway, Maggie Mills & Thomas Poston, *War Reparations: The Case for Countermeasures*, 76 STAN. L. REV. 971 (2024); Lisandra Novo, Specially Affected States' Push for Collective Countermeasures, in 16TH INTERNATIONAL CONFERENCE ON CYBER CONFLICT: OVER THE HORIZON (CYCON) 235, NATO CCDCOE 235 (2024), https://doi.org/10.23919/CyCon62501.2024. 10685582; Jeff Kosseff, *The International Legal Framework for Hunt Forward and the Case for Collective Countermeasures*, in 16TH INTERNATIONAL CONFERENCE ON CYBER CONFLICT: OVER THE HORIZON (CYCON) 221, NATO CCDCOE 221 (2024), https://doi.org/10.23919/CyCon62501.2024.10685559.

of the injured State; (2) necessity for third-party intervention; (3) desired end-state of collective countermeasures; and (4) compliance with the law of countermeasures.

2. COLLECTIVE COUNTERMEASURES IN CYBERSPACE

In *Nicaragua v. U.S.*, the International Court of Justice "could not justify countermeasures taken by a third State… and particularly could not justify intervention involving the use of force."[8] At the time, many legal experts interpreted the opinion to require a "bilateral approach,"[9] barring non-injured States from offering or carrying out any intervention via collective countermeasures. Among the countries that take a bilateral approach to countermeasures are France[10] and Canada.[11] Nevertheless, other major stakeholders do not view Nicaragua as an insurmountable barrier to lawful collective countermeasures.

The 2017 *Tallinn Manual 2.0* considers collective countermeasures "unsettled," while acknowledging the developing view that States not directly injured may have a valid right to respond to breaches of international law or act to protect a collective interest.[12] Additionally, the International Group of Experts did not reach a consensus about the legality of assisting an injured State in carrying out countermeasures, nor on whether countermeasures taken on behalf of an injured State are lawful.[13]

In 2019, Estonia became the first State to endorse collective countermeasures. President Kersti Kaljulaid announced the position that "states which are not directly injured may apply countermeasures to support the state directly affected by the malicious cyber operation,"[14] adopting a "collectivist approach" to countermeasures in cyberspace.[15]

[8] Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.), Judgment, 1986 I.C.J. 14, 249 (June 27) [hereinafter *Nicaragua*].

[9] *Id.*

[10] French Ministry for Europe & Foreign Affairs, *International Law Applied to Operations in Cyberspace* 4 (2021), https://documents.unoda.org/wp-content/uploads/2021/12/French-position-on-international-law-applied-to-cyberspace.pdf [hereinafter *French Position Paper*] ("counter-measures must be taken by France in its capacity as victim. Collective counter-measures are not [authorized], which rules out the possibility of France taking such measures in response to an infringement of another State's rights").

[11] Government of Canada, *International Law Applicable in Cyberspace*, 37 (Apr. 2022), https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/peace_security-paix_securite/cyberspace_law-cyberespace_droit.aspx?lang=eng#a9 [hereinafter *Canadian Position Paper*] ("Canada has considered 'collective cyber countermeasures' but does not, to date, see sufficient State practice or opinio juris to conclude that these are permitted").

[12] TALLINN MANUAL 2.0 ON THE INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS 111, 130–31 (Michael N. Schmitt ed., 2d ed. 2017) [hereinafter TALLINN MANUAL 2.0].

[13] *Id.* at 132.

[14] Kersti Kaljulaid, President of Estonia, Opening Remarks at CyCon 2019 (May 29, 2019), https://president.ee/en/official-duties/speeches/15241-president-of-the-republic-at-the-opening-of-cycon-2019/; *see also* Schmitt & Watts, *supra* note 6.

[15] Schmitt & Watts, *supra* note 6, at 377.

Despite their purported bilateral orientations, the French and Canadian acceptance of certain international obligations brings them closer to the collectivist approach than one might think. Canada maintains that a State owes a due diligence obligation in cyberspace not to permit malicious cyber activities that cause harm to other States.[16] Likewise, France recognizes that the law of neutrality applies in cyberspace, agreeing that a neutral State has an obligation to prevent the use of its territory for malicious purposes.[17] The contemporary development of collective countermeasures indicates that this area of the law may be settling in favor of recognizing the collectivist approach.

A. Academic Support for Collective Countermeasures

Academia has been receptive to the lawfulness of the collectivist approach,[18] emphasizing a utilitarian justification for its application in cyberspace. This practical understanding privileges a global, collective interest in preventing malicious actors in cyberspace from disrupting international order.[19] Sympathetic academics emphasize that *Nicaragua* is not a blanket prohibition on collective countermeasures.

Michael Schmitt and Sean Watts, for example, argue that collective countermeasures in cyberspace are lawful based on the following: (1) there is no prohibition preempting a collectivist approach; (2) there is a practical need for a collaborative mechanism by which to correct asymmetries in cyber capabilities; and (3) the object and purpose of the countermeasures doctrine is consistent with a collectivist the interpretation.[20] Schmitt and Watts' claim that there is no prohibition preempting a collectivist approach[21] challenges the traditional reading of *Nicaragua* as a total ban on collective action.[22] They distinguish *Nicaragua* because it concerned U.S. actions exceeding the use-of-force threshold and there was a lack of injured party consent.[23]

Alternatively, Jeff Kosseff concedes that *Nicaragu*a was a rejection of collective countermeasures appropriate to a kinetic environment. He argues that the interpretation should not be carried over into cyberspace because of the "highly interconnected

[16] *Id*. 26.

[17] *French Position Paper, supra* note 10, at 17.

[18] *See, e.g*., Schmitt & Watts, *supra* note 6; Roguski, *supra* note 7; Kosseff, *supra* note 7; Talita Dias, *Countermeasures in International Law and Their Role in Cyberspace*, CHATHAM HOUSE (2024); Samuli Haataja, *Cyber Operations and Collective Countermeasures Under International Law*, 25 J. CONFLICT & SEC. L. 33 (2020); *but se*e Miles Jackson & Frederica Paddeu, *Proxy Countermeasures in International Law*, EJIL: TALK! (July 5, 2024) ("while the [cyber literature] and practice sometimes use the term 'collective countermeasures' to parallel the notion of 'collective self-defence', the right of collective self-defence cannot provide a compelling analogy… collective self-defence may be rationalised on the basis of the erga omnes character of the prohibition of force").

[19] *See* Schmitt & Watts, *supra* note 6, at 403.

[20] *Id*. at 410.

[21] *Id*. at 403.

[22] *See supra* text accompanying note 8–11.

[23] Schmitt & Watts, *supra* note 6, at 410.

nature of threats in cyberspace,"[24] highlighting how transnational digital networks and the seamless flow of data blur traditional notions of territorial sovereignty.

Oona Hathaway, Maggie Mills, and Thomas Poston echo and cite Schmitt and Watts,[25] adding that the relevant State actions in *Nicaragua* did not violate an international obligation, making the Court's conclusion inapplicable to collective countermeasures in cyberspace where those violations do exist. It is this argument that offers a pathway toward operationalizing collective countermeasures as a concept.

Hathaway, Mills, and Poston argue that "collective countermeasures can be made in response to a state's violation of an obligation *erga omnes*—that is, obligations arising 'towards the international community as a whole' in the protection of which all states have a 'legal interest.'"[26] Drawing upon a "seminal obligation *erga omnes*" to abstain from acts of aggression derived from Article 2 of the United Nations Charter,[27] they argue that the violation of the international obligation allows States other than the injured State to rely on Article 48(1) of the Articles of State Responsibility to "invoke the responsibility of another State" and justify collective countermeasures.[28] In other words, the international obligation to abstain from acts of aggression yields an *erga omnes* obligation in cyberspace to police malicious actors who are conducting attacks from within that States' territory. This perspective is not limited to academia, and recent support from the international community beyond Estonia exhibits growing support for the idea.[29]

B. States Endorsing the Collectivist Approach

Austria,[30] Costa Rica,[31] Ireland,[32] and Poland[33] have all published national position papers that join Estonia in embracing the collectivist approach to countermeasures

[24] *See* Kosseff, *supra* note 7, at 29.

[25] *See* Hathaway, Mills & Poston, *supra* note 7, at 1025.

[26] *Id*. at 1024 (citing Barcelona Traction, Light and Power Company, Ltd. (Belg. v. Spain), Judgment, 1970 I.C.J. 3, 32 (Feb. 5)).

[27] U.N. Charter art. 2, 4.

[28] Hathaway, Mills & Poston, *supra* note 7, at 1024 (citing *Articles on State Responsibility, supra* note 2, art. 48(1)).

[29] *See* Novo, *supra* note 7; *see also* 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/.

[30] Federal Ministry for European and International Affairs of Austria, *Position Paper of the Republic of Austria: Cyber Activities and International Law* (Apr. 2024), https://docslibrary.unoda.org/OpenEnded_Working_Group_on_Information_and_Communication_Technologies__(2021)/Austrian_Position_Paper__Cyber_Activities_and_International_Law_(Final_23.04.2024).pdf [hereinafter *Austrian Position Paper*].

[31] Ministry of Foreign Affairs of Costa Rica, *Costa Rica's Position on the Application of International Law* in Cyberspace (July 21, 2023) https://docs-library.unoda.org/Open-EndedWorkingGrouponInformationandCommunicationTechnologies-(2021)/CostaRica-_Position_Paper_-_International_Law_in_Cyberspace.pdf [hereinafter *Costa Rican Position Paper*].

[32] Department of Foreign Affairs, *Ireland Position Paper on the Application of International Law in Cyberspace* (July 6, 2023) https://www.dfa.ie/media/dfa/ourrolepolicies/internationallaw/Ireland---National-Position-Paper.pdf [hereinafter *Irish Position Paper*].

[33] Ministry of Foreign Affairs Republic of Poland, *The Republic of Poland's Position on the Application of International Law in Cyberspace* (Dec. 29, 2022), https://www.gov.pl/attachment/3203b18b-a83f-4b92-8da2-fa0e3b449131 [hereinafter *Polish Position Paper*].

in cyberspace based on international obligations. The Austrian position cites Article 48(1), invoking State responsibility, and says that collective countermeasures are permitted in instances of violations of *erga omnes* obligations.[34] The Costa Rican[35] and Irish[36] positions mirror this approach, while Poland holds that "the evolution of customary international law over the last two decades provides grounds for [recognizing] that a state may take countermeasures in pursuit of general interest."[37]

Among those supportive, but slightly more cautious, are Denmark, New Zealand, and the United Kingdom. Denmark's position calls the doctrine "unsettled" but nevertheless says there "may be instances" where an injured State can lawfully request assistance applying countermeasures if an international obligation is violated.[38] Similarly, the New Zealand position expresses that it is "open" to injured States requesting assistance with proportional countermeasures,[39] and the United Kingdom has said it believes States are "[open] to consider how the international law framework accommodates, or could accommodate, calls by an injured State for assistance in responding collectively."[40]

Despite developments in favor of the collective approach, there is no unifying standard by which collective countermeasures can be properly evaluated. A practical framework is required to operationalize the concept and create a stable environment in which State practice is governed by strong international norms.

In examining parallels between collective countermeasures and collective self-defense, the goal is to establish a uniform, replicable set of considerations for their use in response to malicious cyber activity. The next section explores and defends the adoption of a normative framework derived from collective self-defense, intended to standardize the practice of collective countermeasures and to invite the cooperation of those who currently favor the bilateral approach.

[34] *Austrian Position Paper, supra* note 30, at 9.

[35] *Costa Rican Position Paper, supra* note 31, at 5 ("countermeasures may be taken by the injured state… a well as third States in response to violations of obligations of an *erga omnes* nature or upon request by the injured State. Thus, States may respond collectively to cyber or non-cyber operations that amount to internationally wrongful acts.").

[36] Ireland stated that "state practice indicates that such measures are permissible in limited circumstances, in particular in the context of violations of peremptory norms." *Irish Position Paper, supra* note 32, 26.

[37] *Polish Position Paper, supra* note 33, at 8.

[38] Jeppe Mejer Kjelgaard & Ulf Melgaard, *Denmark's Position Paper on the Application of International Law in Cyberspace*, 92 NORDIC J. OF INT'L L. 446, 454 (July 4, 2023) https://doi.org/10.1163/15718107-20230001.

[39] New Zealand Ministry of Foreign Affairs & Trade, *The Application of International Law to State Activity in Cyberspace* (Dec. 1, 2020), https://www.mfat.govt.nz/en/media-and-resources/the-application-of-international-law-to-state-activity-in-cyberspace.

[40] Attorney General Suella Braverman, Parliament of the United Kingdom, Speech at Chatham House: International Law in Future Frontiers (May 19, 2022).

3. JUSTIFYING THE APPLICATION OF AN UNWILLING OR UNABLE MODEL

A. The Law of Neutrality and the Due Diligence Principle

Article 2(4) prohibits acts of aggression, but Article 51 permits collective self-defense in response to breaches of an *erga omnes* norm.[41] Collective self-defense refers to a third-party response to attacks by a non-State actor against an injured State conducted because the territorial State is unwilling or unable to neutralize the threat itself. In 2012, Ashley Deeks published a normative framework for applying collective defense.[42] At the time, collective self-defense was an emerging concept, and the lack of international consensus rendered each application reliant on an original set of considerations, effectively allowing auto-interpretation by each State. Deeks relied on doctrinal history and State practice to develop six factors for a replicable "unwilling or unable test" governing third-party responses:

1. Prioritization of consent or cooperation
2. Nature of the threat posed by the non-State actor
3. Request to address the threat and time to respond
4. Reasonable assessment of territorial State control and capacity
5. Proposed means to suppress the threat
6. Prior interactions with the territorial State[43]

In addition to this test, Deeks identified a "historical lineage" of the test in international law, unpacking the legal justification supporting collective self-defense.[44]

The lineage traced collective self-defense to neutrality laws in the context of international armed conflicts between two States. The law of neutrality allows States not party to an armed conflict to demand that their territory not be used as a host for prohibited conduct; bars belligerent States from using neutral territory in furtherance of the conflict; and most importantly requires neutral States to take steps to stop violations of neutrality by belligerent States, should they occur.[45] It is from this law of neutrality that States incur a due diligence obligation to prevent violations of the law of neutrality, which may include use of force on behalf of the neutral State.[46] However, when States are either unwilling or unable to enforce neutrality laws and

[41] Jackson & Paddeu, *supra* note 18.

[42] Ashley S. Deeks, *Unwilling or Unable: Toward a Normative Framework for Extraterritorial Self-Defense*, 52 VA J. INT'L L., 483–550 (2012).

[43] *Id*. at 519–31.

[44] *Id*. at 496.

[45] *Id*. at 497 (citing STEPHEN NEFF, THE RIGHTS AND DUTIES OF NEUTRALS: A GENERAL HISTORY 218 (2000) and NICOLAS POLITIS, NEUTRALITY AND PEACE 21–22 (1935)).

[46] *Id*. at 498 (citing Hague V, art. 10 ("The fact of a neutral Power resisting, even by force, attempts to violate its neutrality cannot be regarded as a hostile act.")).

fail to fulfill their due diligence obligations, collective self-defense has developed into a mechanism for third-party intervention to stop serious threats.[47]

The unwilling or unable test then "migrated into the rules governing a state's use of force extraterritorially against nonstate actors,"[48] gaining international support during the non-international armed conflicts of the late 20th and early 21st centuries.[49] This acceptance of "unwilling or unable" within international law has allowed collective self-defense to counteract kinetic threat actors who might otherwise seek to enact harm on a global stage.

The parallels between collective self-defense and countermeasures are obvious. In a 2019 speech, the Estonian president said, "International security and the rules-based international order have long benefited from collective efforts to stop the violations. We have seen this practice in the form of collective [self-defense] against armed attacks."[50] However, beyond the obvious nomenclature, like collective self-defense, the due diligence obligation also provides a foothold for justifying collective countermeasures in international law.

B. Due Diligence Obligations in Cyberspace

The law of neutrality and subsequent due diligence obligations are widely accepted as applicable to cyberspace.[51] Even States that are silent on or opposed to collective countermeasures, including France, acknowledge that States have an obligation to ensure that their territory[52] is not being used to commit internationally wrongful acts.[53] In 2021, the United States made the following statement regarding the notion of a general obligation of due diligence:

[47] Historical examples supporting the emergence of this doctrine include the Turkish use of force in Iraq in 1996, the Russian use of force in Georgia in 2002, and the United States' intervention in Pakistan in 2007 and 2011. *Id*. at 486–87.

[48] *Id*. at 501.

[49] Elena Chacko & Ashley Deeks, *Which States Support the "Unwilling and Unable" Test?*, LAWFARE (Oct. 10, 2016), https://www.lawfaremedia.org/article/which-states-support-unwilling-and-unable-test.

[50] Kaljulaid, *supra* note 14.

[51] TALLINN MANUAL 2.0 acknowledges due diligence: "a private firm in the first State is engaging in harmful cyber operations in the second State… it would be inappropriate for the second State to launch countermeasures against the firm unless the firm's action can be attributed to the first State… or that State has wrongfully failed to control the activities of the firm and therefore breached its due diligence obligation to control its territory once it became aware of the operations (Rules 6–7)." TALLINN MANUAL 2.0, *supra* note 12, at 113, 6–8.

[52] Generally, a State's "territory" in cyberspace encompasses the physical, digital infrastructure, data and online activities within its jurisdiction.

[53] *See, e.g., Austrian Position Paper, supra* note 30, at 10 ("States are under an obligation to ensure that their territory is not knowingly used for cyber activities contrary to the rights of other states."); *French Position Paper, supra* note 10, at 7–17 ("Under the due diligence obligations, States should ensure that their sovereign domain in cyberspace is not used to commit internationally unlawful acts… The law of neutrality applies to cyberoperations."); Ministry of Foreign Affairs of the People's Republic of China, *China's Views on the Application of the Principle of Sovereignty in Cyberspace* (2021), https://documents.unoda.org/wp-content/uploads/2021/12/Chinese-Position-Paper-on-the-Application-of-the-Principle-of-Sovereignty-ENG.pdf ("No State shall knowingly allow its territory, or territory or ICT facilities, data and information under the control of its government, to be used for ICT activities that undermine national security or interests.").

> The United States has not identified the State practice and *opinio juris* that would support a claim that due diligence currently constitutes a general obligation under international law. We do believe, however, that if a State is notified of harmful activity emanating from its territory it must take reasonable steps to address such activity.[54]

The concern regarding a lack of State practice to establish a rule of customary international law was shared at the time by the United Kingdom[55] and Israel.[56]

Although the United States did not embrace due diligence as an international obligation, other States and international groups have more recently adopted it as a primary rule under customary international law,[57] telegraphing a movement toward broader acceptance.[58] Nevertheless, even as a norm not fully crystallized in customary international law, the due diligence obligation could and should justify collective countermeasures as it does collective self-defense.

Like collective self-defense,[59] the position of Hathaway, Mills, and Poston justifies collective countermeasures under Article 48(1) based on the *erga omnes* obligation to refrain from acts of aggression.[60] Therefore, just as Deeks traces a due diligence obligation through the law of neutrality to justify a third-party response to an *erga omnes* breach in collective self-defense, so too should due diligence in cyberspace be conceptualized as a subset of the prohibition on aggression. In other words, because

[54] Official Compendium of Voluntary National Contributions on the Subject of How International Law Applies to the Use of Information and Communications Technologies by States, U.N. Doc. A/76/136, at 141 (Aug. 10, 2021). One consideration here is that State practice in the realm of cyberspace poses a perpetual hurdle to establishing true international obligations in cyberspace because the speed of technological development has so greatly outpaced the supporting legal infrastructure.

[55] United Kingdom Foreign, Commonwealth & Development Office, *Application of International Law to States' Conduct in Cyberspace: UK Statement* (June 3, 2021).

[56] Roy Schöndorf, *Israel's Perspective on Key Legal and Practical Issues Concerning the Application of International Law to Cyber Operations* 8 (Dec. 2020) [hereinafter Israeli Position Paper].

[57] *See Irish Position Paper, supra* note 32, 12–13. ("the due diligence principle [is] a primary rule of international law. Therefore, a breach of this international obligation, which is attributable to a state, engages state responsibility"); *see also* African Union Peace and Security Council, *Common African Position on the Application of International Law to the Use of Information and Communication Technologies in Cyberspace* 21 (Jan. 29, 2024) [hereinafter *African Union Position Paper*] ("due diligence is an obligation that operates in the context of other primary rules of international law. … every State is under an obligation"); *Costa Rican Position Paper, supra* note 31, 27 ("Under customary international law, States have a general obligation 'not to allow knowingly its territory to be used for acts contrary to the rights of other States'."); *French Position Paper, supra* note 10, at 6 ("the due diligence requirement… is a customary obligation for States, which must… ensure that their territory is not used for [wrongful acts using ICTs] including by non-state actors").

[58] Ashley Deeks, *Defend Forward and Cyber Countermeasures, Hoover Working Group on National Security, Technology, and Law*, in AEGIS SERIES PAPER NO. 2004, at 9 (Aug. 4, 2020) (quoting ANN VÄLJATAGA, TRACING *OPINIO JURIS* IN NATIONAL CYBER SECURITY STRATEGY DOCUMENTS 15 (2018)) ("A researcher who reviewed recent statements by Western states summarized those statements as reflecting a shift in emphasis from self-defense to countermeasures, a 'general approval of collective response,' and a sense that the *opinio juris* in national strategies 'is currently bent towards overriding the prohibition on collective countermeasures'").

[59] *See supra* text accompanying notes 42–48.

[60] *See* Hathaway, Mills & Poston, *supra* note 7, at 1027.

a State is prohibited *erga omnes* from acts of aggression, they incur an obligation of conduct[61] to exercise "due diligence" in addressing threats emitting from their territory. Therefore, where a territorial State fails to repel the bad actor, Article 48(1) allows a third-party State to step in and take collective countermeasures to assist the injured State and induce compliance with international law.[62]

4. THE UNWILLING OR UNABLE PARADIGM IN COUNTERMEASURES

Despite growing support for collective countermeasures, legal justifications will only serve the collective interest insofar as there exists a unified standard by which to evaluate and apply them in cyberspace. Under this paradigm of due diligence as an obligation derived from the *erga omnes* prohibition on aggression, the unwilling or unable test is a natural benchmark for operationalizing the concept. Unwilling or unable provides a model threshold for one State to use another's inaction, whether from incapacity or indifference, as justification for a third-party response. This paper proposes a four-factor test for assisting States to evaluate the lawfulness of launching—or assisting in the launch of—countermeasures on behalf of an injured State.

A. Consent and Cooperation of the Injured State

Practical interest in collective countermeasures lies primarily in curing vulnerabilities from asymmetrical cyber infrastructure across different States and inhibiting those who might overclassify events to evoke collective self-defense rights. Injured State consent is the most important condition for the right to launch countermeasures on behalf of another State.

Where a State is unable to respond itself and consents to third-party action, there is a strong case for the use of collective countermeasures. Conversely, the denial of consent by an injured State carries extreme weight in this analysis. Where a technologically capable injured State is unwilling to respond itself, only extreme necessity would overcome the lack of consent.

Unlike in collective self-defense, consent cannot negate the unwilling or unable analysis in the context of collective countermeasures. Consent is a crucial aspect of distinguishing the prohibition on collective action in *Nicaragua* from the lawful collective countermeasures contemplated by this paper.[63] Rather, consent is a recurring consideration of the State positions and a very strong factor in support of

[61] For an in-depth evaluation of obligations of conduct versus obligations of result, see Alice Ollino, *The Nature of Due Diligence Obligations*, in DUE DILIGENCE OBLIGATIONS IN INTERNATIONAL LAW 64, 64–130 (Cambridge Univ. Press, 2022).

[62] *See supra* text accompanying note 28.

[63] *See supra* text accompanying note 23.

collective countermeasures.[64] It remains an open question whether the violation of an erga omnes norm alone would be enough to distinguish the prohibition in *Nicaragua*.

B. Necessity for Third-Party Intervention

This prong evaluates the risk to the international community and will identify the specific interests at stake in each case. Broadly, this test encompasses projected political and diplomatic ramifications, disruptions to businesses and trade, economic consequences, international norms, and other cooperative challenges. More specifically, this assessment includes the technical risks and nature of the threat. The greater the threat to the international community, the more necessary third-party action becomes.

Technical risks are necessarily fact-dependent but would consider the pattern/severity/frequency of activity and the control and capabilities of the parties involved. These factors could include the potential for cascading effects, exfiltration of sensitive or partner-nation data, interruption of international supply chains, and malicious exploitation of national or third-party software. This analysis would also evaluate the scale and sophistication of past activity, the threat posture of the responsible party, the imminence of further activity, and, ultimately, the vulnerability of the injured State. The more exaggerated the discrepancy between the sophistication of the responsible party and the vulnerability of the injured State, the stronger the case for third-party intervention.

C. Desired End-State of Collective Countermeasures

Countermeasures must only be used for the purpose of inducing the responsible party to return to a state of compliance with international obligations.[65] As reflected in Table I, one must identify the responsible party, one must consider whether the issue concerns malicious cyber activity attributable[66] to a State actor or non-attributable activity by a non-State actor. Once attribution and consent have been assessed, one must identify the desired end-state of the countermeasures based on the posture of the parties involved. The appropriate target and end-state differ depending on whether the activity is attributable or not:

1) Attributable Activity

In the case of attributable cyber activity, a consenting injured State permits the third party to take collective countermeasures against the responsible State. The desired end-state remains constant: to induce cessation of the malicious cyber activity.

[64] *See, e.g.*, *Irish Position Paper, supra* note 32 at 26 ("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").

[65] TALLINN MANUAL 2.0, *supra* note 12, at 112, 5.

[66] This discussion of "attributable activity" refers specifically to activity for which a State is responsible.

2) Non-Attributable Activity

If the malicious cyber activity is performed by a non-State actor, then one must consider whether the territorial State is unable or simply unwilling to address the malicious activity. If the territorial State is unwilling to act (but technologically capable of fulfilling its due diligence obligation), then the assisting State would direct collective countermeasures against the territorial State to incentivize compliance with its due diligence obligations.

If the territorial State is unable to act—whether due to a lack of technological capability or other reasons—then the State would direct countermeasures against the non-State actor itself because the territorial State would be in breach of its due diligence obligation to control its territory.[67] Launching countermeasures against non-State actors remains a controversial application of the doctrine and would require a strong justification under the necessity prong establishing a requirement for collective action.

TABLE I: DESIRED END-STATES OF COLLECTIVE COUNTERMEASURES

<table>
<tr><td rowspan="2"><b>Attributable malicious cyber activity</b></td><td colspan="2"><b>Injured State Unable</b></td></tr>
<tr><td colspan="2">Launch collective countermeasures against responsible State to induce cessation of malicious activity.</td></tr>
<tr><td rowspan="2"><b>Non-attributable malicious cyber activity</b></td><td><b>Territorial State Unwilling</b></td><td><b>Territorial State Unable</b></td></tr>
<tr><td>Launch collective countermeasures against territorial State to induce compliance with due diligence obligations.</td><td>Launch collective countermeasures against non-State actor directly to induce cessation of malicious activity.</td></tr>
</table>

D. Compliance with the Law of Countermeasures

Finally, any countermeasures launched must comply with international legal requirements. Among these requirements, they must respond to a prior breach of international law; they must target the responsible party; they must comply with the proportionality principle; they must not involve the use of force; they must be reversible once the responsible party resumes compliance with international obligations.[68]

[67] TALLINN MANUAL 2.0, *supra* note 12, at 113, 6–8 ("assume that a private firm in the first State is engaging in harmful cyber operations in the second State. In such a case, it would be inappropriate for the second State to launch countermeasures against the firm unless the firm's action can be attributed to the first State… or that State has wrongfully failed to control the activities of the firm and therefore breached its due diligence obligation to control its territory"); *see also* Ollino, *supra* note 61, at 67 ("breaches of *erga omnes* entitle states other than the injured one to invoke the conditions provided by Article 48(2)").

[68] TALLINN MANUAL 2.0, *supra* note 12, at 111–34.

Whether there is a notification requirement prior to taking countermeasures remains disputed in the context of cyberspace.[69]

5. CONCLUSION

With the vast disparities in cyber capabilities across different States, perverse incentives in current practice, and technological development outpacing legal innovation, State practice and *opinio juris* continue to trail. This delta poses a significant barrier to forming a responsive body of law in cyberspace. With a four-factor test considering (1) consent and cooperation of the injured State; (2) necessity for third-party intervention; (3) desired end-state of collective countermeasures; and (4) compliance with the law of countermeasures, collective countermeasures can be consistently executed under international law.

Granted, some questions remain to be explored. For instance, although justifiable via a due diligence rationale, will countermeasures taken directly against a non-State actor be palatable to a majority of States? Additionally, States' interpretations of international law vary significantly, injecting uncertainty into the collective countermeasures doctrine in application. Finally, States' interpretation of their sovereign cyberspace, especially given the role of multinational technology companies, is convoluted, creating potential misalignments in what constitutes a State's territory.

These questions notwithstanding, collective countermeasures provide the international community with a crucial tool for policing malicious cyber activity and maintaining a peaceful international network. Establishing a threshold for joint action against hostile activity via collective countermeasures would result in a more secure and organized global network. This test is a starting point for a principled and consistently replicable framework for the collective defense and safeguarding of cyberspace. Given a strong legal backing, the hope remains that with future development, establishing a unified standard for collective action allows State practice and opinio juris to mature, potentially garnering further support for the collectivist approach.

[69] *Id*. at 120, 10–12.

CyCon 2025: The Next Step
17th International Conference on Cyber Conflict
C. Kwan, N. Gratzer, K. Podiņš, M. Tolppa (Eds.)
2025 © NATO CCDCOE Publications, Tallinn

Permission to make digital or hard copies of this publication for internal use within NATO and for personal or educational use when for non-profit or non-commercial purposes is granted provided that copies bear this notice and a full citation on the first page. Any other reproduction or transmission requires prior written permission by NATO CCDCOE.

The Next Step in Global Connectivity: Legal Challenges in the Shift from Subsea Cables to Satellites

Anna Blechová
PhD Student
Institute of Law and Technology
Faculty of Law
Masaryk University
Brno, Czech Republic
anna.blechova@law.muni.cz

Abstract: In today's digitalized society, our daily lives are inextricably linked to cyberspace and the technologies that sustain it. Thus, the protection of critical infrastructure, such as internet infrastructure, has become a priority. However, ongoing international conflicts, rising political tensions, and the increasing likelihood of human error – capable of causing global cyber outages – are forcing a re-evaluation of our previous decisions in this domain.

Subsea fibre-optic cables, responsible for carrying more than 95% of international data, have emerged as high-risk targets for cyber operations and potential threats of power struggles between states like the United States, China, and Russia. In response, initiatives such as NATO's Science for Peace and Security Programme or those coming from Taiwan are exploring the next step or 'Plan B' – the development of a more resilient global internet infrastructure built on secure satellite networks.

This paper investigates the legal challenges facing both subsea cables and satellite infrastructure as critical components of global connectivity. While satellite infrastructure may initially appear more resilient, this paper argues that the existing regulatory framework and current geopolitical landscape could undermine its perceived advantages and that legislation for the sea is more evolved than that for outer space. Moreover, in the context of armed conflict, reliance on satellite networks may introduce vulnerabilities that could generate even greater uncertainties than those posed by subsea cables. The findings highlight the need for a clearer legal and regulatory approach to secure both subsea and satellite infrastructure, which will make explicit the rules of responsibility and liability in these arenas, in the evolving landscape of cyber warfare.

Keywords: *space law, satellites, subsea cables, cybersecurity, law of the sea, responsibility, New Space*

1. INTRODUCTION

This article owes its existence to subsea cables. Indeed, if you are accessing this paper online, you are directly benefitting from the indispensable role these conduits play in global communication. Despite being no wider than a common garden hose, subsea cables constitute the backbone of modern connectivity, with an estimated 95% of global internet traffic traversing these systems. It is estimated that the data for more than USD 10 trillion in financial transactions is transmitted every day via subsea cables.[1] Therefore, David Cattler, NATO's assistant secretary general for intelligence and security, has called subsea cables the linchpin of the modern information economy.[2] However, this reliance may evolve, and in the future, access to this paper could increasingly depend on satellite infrastructure. Given that satellite networks are considered a viable future alternative to subsea cables, this article aims to compare the legal frameworks governing both domains and explore the similarities and challenges they present.

An estimated 500 to 600 subsea cables span the world's oceans, extending about 1.2 million kilometres in all.[3] If enough of these cables were damaged, it could profoundly disrupt global communication and daily life.[4] While the failure of a single cable might have limited immediate consequences due to redundancy mechanisms, scenarios involving simultaneous damage to multiple cables, compounded by the limited availability of repair vessels, could trigger cascading connectivity failures.[5]

[1] Tim Stronge, 'Do $10 Trillion of Financial Transactions Flow Over Submarine Cables Each Day?' (*TeleGeography*, 6 April 2023) <https://blog.telegeography.com/2023-mythbusting-part-1> accessed 9 January 2025.

[2] Charlie Cooper, 'NATO Warns Russia Could Target Undersea Pipelines and Cables' (*Politico*, 3 May 2023) <https://www.politico.eu/article/nato-warns-russia-could-target-undersea-pipelines-and-cables/> accessed 13 January 2025.

[3] Edmon de Haro, 'NATO Plans an Orbital Backup Internet Using Satellite Broadband' (*IEEE Spectrum*, 24 December 2024) <https://spectrum.ieee.org/undersea-internet-cables-nato> accessed 9 January 2025.

[4] Kamal Acharya, 'A Sinking Ship and the Fragility of the Internet: How NATO Plans to Secure the World's Digital…' (*Medium*, 1 January 2025) <https://medium.com/@lotussavy/a-sinking-ship-and-the-fragility-of-the-internet-how-nato-plans-to-secure-the-worlds-digital-bdc348c6b56f> accessed 9 January 2025.

[5] Douglas Burnett, 'Submarine Cable Security and International Law' (2021) 97 *International Law Studies* 1661–1663 <https://digital-commons.usnwc.edu/ils/vol97/iss1/55> accessed 9 January 2025.

This potential vulnerability underscores the critical importance of robust protective measures. Moreover, key islands, such as Iceland, Cyprus, Malta, and Ireland, play an essential role.[6] Iceland hosts several data centres supporting financial services and cloud computing, and it is connected to the global network through four submarine cables.[7] An attack on these cables would create challenges for Iceland and have significant implications for Europe and the United States.

Recent incidents have highlighted the increasing frequency and severity of threats to subsea cables, both physical and cyber in nature. Amid the current geopolitical climate, these threats have emerged as a major concern, as demonstrated by a number of high-profile attacks and their wide-ranging implications.[8] For instance, damage to cables in the Malacca Strait and Java Sea exemplifies vulnerabilities in regions with high traffic and limited backup systems.[9] The Baltic Sea has emerged as a critical focal point in the context of global security, particularly against the backdrop of the war in Ukraine. From the onset of the conflict, Russia has been actively mapping subsea cables in the region, raising concerns about potential vulnerabilities.[10] Simultaneously, incidents involving Chinese vessels have added to the tension. In October 2023, the *NewNew Polar Bear*, a Hong Kong-flagged, Chinese-registered ship, allegedly damaged two subsea data cables and a gas pipeline in the Baltic Sea. A second incident occurred in November 2024, when the Yi Peng 3, a Chinese cargo vessel, reportedly severed two communications cables connecting Germany to Finland and Lithuania to Sweden.[11] These events underscore the growing risks to vital undersea infrastructure in the region.[12] In June 2022, damage to a cable in Egypt triggered a significant internet outage across seven countries. Ethiopia experienced a 90% loss in connectivity, while Somalia faced an 85% reduction. The disruption also affected cloud services provided by Google, Amazon, and Microsoft.[13]

[6] S Besch and E Brown, 'Securing Europe's Subsea Data Cables' (*Carnegie Endowment for International Peace*, 16 December 2024) 5 <https://carnegieendowment.org/research/2024/12/securing-europes-subsea-data-cables?lang=en> accessed 13 January 2025.

[7] de Haro (n 3); Tom Porter, 'NATO Is Working to Reroute Data through Space, Fearing Russia Could Slice Undersea Internet Cables' (*Business Insider*, 2 January 2025) <https://www.businessinsider.com/nato-plan-to-defend-undersea-internet-from-sabotage-using-satellites-2025-1> accessed 9 January 2025.

[8] Besch and Brown, 'Securing Europe's Subsea Data Cables' (n 6).

[9] Elina Noor, 'Subsea Communication Cables in Southeast Asia: A Comprehensive Approach Is Needed'(*Carnegie Endowment for International Peace*, 18 December 2024) <https://carnegieendowment.org/research/2024/12/southeast-asia-undersea-subsea-cables?lang=en&center=russia-eurasia> accessed 13 January 2025.

[10] Jim Sciutto, 'Exclusive: US Sees Increasing Risk of Russian "Sabotage" of Key Undersea Cables by Secretive Military Unit' (*CNN*, 6 September 2024) <https://www.cnn.com/2024/09/06/politics/us-sees-increasing-risk-of-russian-sabotage-undersea-cables/index.html> accessed 9 January 2025; S Besch and E Brown, 'A Chinese-Flagged Ship Cut Baltic Sea Internet Cables. This Time, Europe Was More Prepared.' (*Carnegie Endowment for International Peace*, 3 December 2024) <https://carnegieendowment.org/emissary/2024/12/baltic-sea-internet-cable-cut-europe-nato-security?lang=en> accessed 9 January 2025.

[11] H Astier and P Kirby, 'Germany Suspects Sabotage over Severed Undersea Cables in Baltic' (*BBC*, 19 November 2024) <https://www.bbc.com/news/articles/c9dl4vxw501o> accessed 9 January 2025; Besch and Brown, 'Securing Europe's Subsea Data Cables' (n 6).

[12] Besch and Brown, 'A Chinese-Flagged Ship Cut Baltic Sea Internet Cables' (n 10).

[13] 'The Most Vulnerable Place on the Internet' (*WIRED*) <https://www.wired.com/story/submarine-internet-cables-egypt/> accessed 9 January 2025.

Another example is the incident in the Red Sea. In February 2024, a missile strike by Yemen's Houthi militants targeted the cargo ship *Rubymar* in the Red Sea. Following the crew's evacuation, the vessel, left adrift, severed three major undersea fibre-optic cables, responsible for transmitting a quarter of the Internet traffic between Europe and Asia. This incident disrupted global data networks, highlighting the vulnerability of critical digital infrastructure.[14] This could suggest that attacks on subsea cables are limited to the Western Hemisphere, but that is not the case. For instance, in 2023, the subsea cable connecting Taiwan and the Matsu Islands, located near the Chinese coast, was severed, cutting off internet access for the 14,000 residents. Taiwanese authorities speculated that China could be responsible, although no evidence supported this claim. Given the recurrence of such incidents, Taiwan has recognized the vulnerability of its connections to the global network and has begun exploring solutions, including rerouting internet infrastructure via outer space.[15]

While previous examples have primarily addressed physical attacks on subsea cables, cyberattacks also represent a significant threat in this domain.[16] These attacks could be particularly relevant in the context of armed conflict, as subsea cables are integral to military operations and communications,[17] as well as during peacetime, because subsea cables are vital for our modern society. Furthermore, the US Office of the Director of National Intelligence has classified the potential for cyberattacks targeting cable landing stations as a 'high risk' to national security.[18]

This intensifying focus on subsea cable security has emerged not in isolation but rather in response to shifting geopolitical dynamics. Increased tensions among states, particularly between Russia,[19] China, the member states of the European Union, and the United States, have elevated the strategic significance of subsea cables. For example, despite the limited number of subsea cable providers and operators, Europe[20] and the United States[21] have chosen to restrict one of the largest, Huawei Marine

[14] Acharya (n 4).

[15] Sarah Wu and others, 'Fear of the Dark: Taiwan Sees Wartime Frailty in Communication Links with World' (*Reuters*, 16 March 2023) <https://www.reuters.com/world/asia-pacific/fear-dark-taiwan-sees-wartime-frailty-communication-links-with-world-2023-03-15/> accessed 11 January 2025.

[16] Naveen Goud, 'Cyber Threat to Submarine Cables in China Sea - Cybersecurity Insiders' (10 April 2025) <https://www.cybersecurity-insiders.com/cyber-threat-to-submarine-cables-in-china-sea/, https://www.cybersecurity-insiders.com/cyber-threat-to-submarine-cables-in-china-sea/> accessed 14 April 2025.

[17] Besch and Brown, 'Securing Europe's Subsea Data Cables' (n 6).

[18] Andrea Ratiu, 'Cyber Defense across the Ocean Floor: The Geopolitics of Submarine Cable Security' (*Atlantic Council*, 13 September 2021) <https://www.atlanticcouncil.org/in-depth-research-reports/report/cyber-defense-across-the-ocean-floor-the-geopolitics-of-submarine-cable-security/> accessed 1 March 2025.

[19] Peter Dickinson, 'Concerns Grow over Possible Russian Sabotage of Undersea Cables' (*Atlantic Council*, 12 September 2024) <https://www.atlanticcouncil.org/blogs/ukrainealert/concerns-grow-over-possible-russian-sabotage-of-undersea-cables/> accessed 9 January 2025.

[20] 'Texts Adopted - Security and Defence Implications of China's Influence on Critical Infrastructure in the European Union - Wednesday, 17 January 2024' (*European Parliament*) <https://www.europarl.europa.eu/doceo/document/TA-9-2024-0028_EN.html> accessed 11 January 2025.

[21] Joe Brock, 'U.S. and China Wage War Beneath the Waves – Over Internet Cables' (*Reuters*, 24 March 2023) <https://www.reuters.com/investigates/special-report/us-china-tech-cables/> accessed 11 January 2025.

Networks, based in China. This decision reflects growing concerns that Chinese technology may heighten cybersecurity and espionage risks. Moreover, Russia's aggressive posture following its invasion of Ukraine has led to heightened concerns over its potential targeting of critical infrastructure, including subsea cables. Similarly, China's growing interest in leveraging physical, cyber and space-related[22] capabilities to assert influence has also raised alarm.[23] These developments have driven states and international organizations to reassess their approaches to subsea cable protection.

In parallel, the discussion has extended to exploring alternative connectivity solutions, particularly satellite-based networks. Taiwan[24] and NATO,[25] for instance, have launched initiatives to strengthen satellite infrastructure as potential complements or substitutes for subsea cables. While satellite networks offer certain resilience advantages, the growing reliance on them introduces distinct vulnerabilities, especially in the context of armed conflict.

Recent initiatives aimed at enhancing subsea cable security reflect this evolving priority. The United States and the European Union have focused on the protection of subsea cables. Moreover, NATO, besides establishing a Maritime Centre for the Security of Critical Undersea Infrastructure[26] and a Critical Undersea Infrastructure Coordination Cell,[27] has launched the Hybrid Space-Submarine Architecture Ensuring Infosec of Telecommunications (HEIST)[28] programme to safeguard global connectivity. The HEIST project has two primary objectives: first, to rapidly identify the location of damaged submarine cables, and second, to enhance the number of available pathways for data transmission. Specifically, the project will focus on exploring methods to reroute high-priority traffic through satellites in orbit.[29] As Gregory Falco points out, the range of internet pathways should include 'something in the sky rather than [just] what's on the seabed.'[30] This statement is undeniably compelling. While this article aims to address potential challenges, the author firmly believes that diversity in cybersecurity is paramount.

[22] DK Tatlow, 'China's Push for Supremacy Moves into Space' (*Newsweek*, 18 December 2024) <https://www.newsweek.com/2025/01/17/china-space-infrastructure-us-latin-america-chile-argentina-1999644.html> accessed 9 January 2025.

[23] Astier and Kirby (n 11).

[24] J McGillis and P van Wingerden, 'Why Taiwan Needs to Secure Its Undersea Cables' (*Diplomat*, 1 July 2024) <https://thediplomat.com/2024/07/why-taiwan-needs-to-secure-its-undersea-cables/> accessed 9 January 2025.

[25] de Haro (n 3).

[26] NATO, 'Vilnius Summit Communiqué Issued by NATO Heads of State and Government (2023)' (NATO) <https://www.nato.int/cps/en/natohq/official_texts_217320.htm> accessed 11 January 2025.

[27] 'NATO Stands up Undersea Infrastructure Coordination Cell' (*NATO*) <https://www.nato.int/cps/en/natohq/news_211919.htm> accessed 11 January 2025.

[28] 'Home' (*Heist a NATO SPS Project*)<https://natoheist.org/Home.html> accessed 9 January 2025; de Haro (n 3).

[29] de Haro (n 3).

[30] ibid.

The primary objective of this article is to investigate the legal challenges associated with the protection of subsea cables and satellite infrastructure as critical components of global connectivity. The analysis is structured into three sections. Section 2 of the paper examines the legal status of subsea cables in the context of cyberattacks. Section 3 evaluates the legal framework governing satellite infrastructure under analogous conditions. Section 4 provides a comparative analysis of both systems, describing their respective strengths and weaknesses.

While satellite infrastructure may initially appear more resilient, this paper argues that the current regulatory and geopolitical landscape may undermine these perceived advantages. Moreover, in the context of armed conflict, reliance on satellite networks could introduce vulnerabilities even more significant than those associated with subsea cables. Thus, the findings emphasize the urgent need for a comprehensive legal and regulatory framework capable of safeguarding both subsea and satellite infrastructure in the evolving landscape of cyber warfare, underscoring the importance of carefully tailored approaches to address these dual challenges.

2. UNDER THE SEA OR UNDER THE STARS: LEGAL REGIMES WITHOUT BOUNDARIES

This section examines the legal framework governing subsea cables and space infrastructure, focusing on their vulnerability to cyberattacks and other forms of hostile interference. The analysis will commence with an in-depth exploration of the legal status and regulatory landscape surrounding subsea cables. Then the discussion will shift to an equally detailed assessment of the legal status of space assets, encompassing satellites and related infrastructure, which are increasingly integral to modern technological and security frameworks. This approach seeks to elucidate the unique challenges and legal implications associated with safeguarding these essential systems in the context of contemporary cybersecurity threats. Consequently, it is crucial to examine responsibility regimes to determine who can be held accountable for such activities, whether they occur beneath the sea or among the stars. The attribution of malicious activities has been a topic of considerable discussion among cybersecurity experts, and thus it will also be included in this text.[31]

A. Legal Status of Subsea Cables in Regard to Cyberattacks

The legal status of subsea cables under international law is governed by a combination of treaties and frameworks that, while providing foundational principles, exhibit notable gaps and challenges, particularly in addressing contemporary cybersecurity

[31] Jakub Vostoupal, ‘Stuxnet vs WannaCry and Albania: Cyber-Attribution on Trial’ (2024) 54 *Computer Law & Security Review* 106008 <https://www.sciencedirect.com/science/article/abs/pii/S026736492400075X>; Jason Healey, ‘Beyond Attribution: Seeking National Responsibility for Cyber Attacks’ (*Atlantic Council*, 2012) <https://www.atlanticcouncil.org/wp-content/uploads/2012/02/022212_ACUS_NatlResponsibilityCyber.PDF> accessed 9 January 2025.

concerns. The primary legal instruments include the Convention for the Protection of Submarine Cables,[32] which has been applicable since 1884,[33] the United Nations Convention on the Law of the Sea (UNCLOS),[34] and other agreements such as the Convention on the Continental Shelf[35] or the Geneva Convention on the High Seas.[36] These treaties establish key provisions, such as the protection of subsea cables and mechanisms for cooperation among states. However, they have faced criticism from various scholars for their limitations, particularly in the context of subsea cables' increasing vulnerability to cyber threats.[37] Nevertheless, although UNCLOS is not the only treaty governing this domain, it remains the most widely accepted and enforced. Its status as a universal convention underscores its foundational role in the legal framework regulating subsea infrastructure.

One pressing issue is the lack of attention historically paid to the cybersecurity and physical vulnerabilities of subsea cables. As Davenport has written, national and international security strategies had hitherto neglected to recognize the potential for subsea cables to become a target for malicious activities.[38] Over the past decade, however, awareness has increased. Countries such as France, Germany, and Estonia, alongside international organizations like NATO and the European Union, have now acknowledged the strategic importance of subsea cables and the risks associated with their vulnerability.[39]

Nevertheless, significant gaps remain in the current legal and regulatory framework.[40] For example, under Article 113 of UNCLOS, states are required to establish enforcement mechanisms to hold accountable those who damage subsea infrastructure, whether intentionally or unintentionally, particularly in cases involving vessels flying their flag.[41] Yet, as noted by Davenport and Besch and Brown, a lack of willingness among certain states to implement or enforce such measures has thus far rendered

[32] 'Convention for the Protection of Submarine Telegraph Cables' (1884) <https://web.archive.org/web/20160303182819/http://cil.nus.edu.sg/wp/wp-content/uploads/2009/10/Convention_on_Protection_of_Cables_1884.pdf> accessed 13 January 2025.

[33] Mikaela Cardillo, 'Navigating International Law Safeguards for Submarine Cables: Charting a Course for Effective Protections' (*SSRN*, 25 September 2023) 317 <https://papers.ssrn.com/abstract=5044559> accessed 9 January 2025.

[34] United Nations Convention on the Law of the Sea (UNCLOS) <https://www.un.org/depts/los/convention_agreements/texts/unclos/unclos_e.pdf> accessed 9 January 2025.

[35] Convention on the Continental Shelf <https://treaties.un.org/pages/viewdetails.aspx?src=treaty&mtdsg_no=xxi-4&chapter=21&clang=_en> accessed 13 January 2025.

[36] Convention on the High Seas <https://treaties.un.org/pages/viewdetails.aspx?src=treaty&mtdsg_no=xxi-2&chapter=21> accessed 13 January 2025.

[37] For example, Cardillo (n 33); LR Wrathall, 'The Vulnerability of Subsea Infrastructure to Underwater Attack: Legal Shortcomings and the Way Forward' (2010) 12 San Diego International Law Journal 223 <https://digital.sandiego.edu/ilj/vol12/iss1/8>; Burnett (n 5); Tara Davenport, 'Submarine Cables, Cybersecurity and International Law: An Intersectional Analysis' (2015) 24 Catholic University Journal of Law and Technology 57–59.

[38] Davenport (n 37).

[39] Besch and Brown, 'A Chinese-Flagged Ship Cut Baltic Sea Internet Cables' (n 10).

[40] Cardillo (n 33) 314.

[41] Davenport (n 37) 83.

these provisions ineffective in practice.[42] This reluctance undermines the protective framework established by UNCLOS and raises concerns about the adequacy of existing international law to address the realities of modern threats.

Given that subsea cables constitute the backbone of global communication, any disruption – whether cyber or physical – could potentially have catastrophic implications for data transmission and connectivity worldwide. In practice, we can observe that amid the escalating trade conflict between the United States and China in 2025, threats to international communication are being used as a tool in geopolitical power play.[43] It is therefore imperative to address these deficiencies, both in terms of legal enforcement and in the development of more robust international standards.

Responsibility and Attribution

The legal framework governing subsea cables highlights distinct differences in the responsibility and attribution regimes depending on their location. These differences stem from the varying legal regimes applicable to territorial waters, exclusive economic zones (EEZs), and the high seas. Territorial water, extending up to 12 nautical miles from a state's baselines,[44] refers to areas where states exercise sovereignty. Within this zone, states can protect their subsea cables and establish and enforce preventive cybersecurity measures. The EEZs, spanning up to 200 nautical miles, provide states with sovereign rights over natural resources and jurisdiction for specific purposes, including environmental protection and cable-related activities.[45] However, these rights are more limited than the full sovereignty exercised within territorial waters. By contrast, the high seas, like outer space, fall outside the jurisdiction of any single state and are not subject to sovereignty under international law. Subsea cables, unlike ships and vessels, are not flagged to any state, and thus assigning the cable a 'nationality' could be challenging.[46] Another problematic aspect is that, for example, EU officials have concerns regarding the ownership of subsea cables and regulatory regimes.[47]

Under international law, states are not automatically responsible for every action of an individual in their jurisdiction.[48] Nevertheless, under specific conditions, states can be held responsible for the actions of individuals even outside of their territory via the quasi-territorial jurisdiction regimes such as the law of the flag. This principle

[42] ibid 83–85; Besch and Brown, 'Securing Europe's Subsea Data Cables' (n 6) 6.

[43] EL Murphy and M Pearl, 'China's Underwater Power Play: The PRC's New Subsea Cable-Cutting Ship Spooks International Security Experts' <https://www.csis.org/analysis/chinas-underwater-power-play-prcs-new-subsea-cable-cutting-ship-spooks-international> accessed 14 April 2025.

[44] 'The 12 Nautical Mile Rule & Its Impact on Maritime Laws' (*Lorrendraaier*, 24 May 2023) <https://lorrendraaier.nl/general/the-12-nautical-mile-rule-its-impact-on-maritime-laws/> accessed 13 January 2025.

[45] Cardillo (n 33) 317–318.

[46] Burnett (n 5).

[47] Besch and Brown, 'Securing Europe's Subsea Data Cables' (n 6) 13.

[48] United Nations and James Crawford (eds), *The International Law Commission's Articles on State Responsibility: Introduction, Text, and Commentaries* (CUP 2002) 91.

also applies to activities under the sea and could align with the Articles on the Responsibility of States for Internationally Wrongful Acts (ARSIWA). Key provisions, such as Articles 1, 2, 28, and 30, establish the criteria for state responsibility and the consequences of internationally wrongful acts. These principles are particularly relevant in the context of subsea cables and share parallels with the legal framework for outer space activities. Nevertheless, it is important to highlight that although ARSIWA is generally accepted and applied, it does not constitute a set of binding rules.

An important phenomenon that complicates the legal landscape is the nature of subsea cables themselves. These cables are lengthy infrastructure systems predominantly owned and operated by private entities, often traversing multiple jurisdictions. As a result, a mosaic of legal regimes may apply, creating significant challenges for legal certainty. Each state could potentially apply different legal standards to the same cable, undermining coherence and predictability in governance. Additionally, since private companies are generally not subject to direct responsibility or liability under international law, this fragmented regime introduces further complications. For instance, addressing cybersecurity threats or attributing cyberattacks on subsea cables becomes exceedingly difficult in such a complex legal environment.

Given the critical importance of subsea cables to global communications and economic stability, there is a pressing need to enhance international cooperation and establish a more harmonized legal framework. This would not only bolster legal certainty but also ensure more effective protection against emerging threats, particularly in the realm of cybersecurity.

B. Legal Status of Satellite Infrastructure in Regard to Cyberattacks

To properly contextualize the legal and regulatory issues surrounding outer space, it is essential first to describe the domain itself. Outer space, particularly concerning satellites and other space assets, has undergone significant transformation in recent years. The emergence of new space actors is turning this domain from one previously dominated by state-led initiatives to one increasingly shaped by private companies.[49] This phenomenon, often referred to as New Space, marks a paradigm shift that fundamentally alters the dynamics of power and authority within the space arena while simultaneously introducing novel legal and regulatory challenges.[50]

Satellite infrastructure, a critical component of outer space activities, is regulated by a set of international agreements collectively known as the ‘cosmic treaties’. These

[49] Arianna Vettorel, ‘Cybersecurity in New Space and the Problem of International Regulation’ (2024) 49 Air and Space Law <https://kluwerlawonline.com/api/Product/CitationPDFURL?file=Journals\AILA\AILA2024025.pdf> accessed 11 January 2025; DP Fidler, ‘Cybersecurity and the New Era of Space Activities’ (*Council on Foreign Relations*, April 2018) <https://www.cfr.org/report/cybersecurity-and-new-era-space-activities> accessed 11 January 2025; Nayef Al-Rodhan, ‘The New Space Race’ [2018] The National Interest 67.

[50] Al-Rodhan (n 49); Gregory Falco, *The Vacuum of Space Cyber Security* (2018).

include, for example, the Outer Space Treaty (OST), the Registration Convention, and the Liability Convention. These treaties establish foundational principles for the governance of outer space, such as the prohibition of national sovereignty, the peaceful use of space, and the allocation of liability for damages. However, despite the rapid technological and commercial advances of the New Space era, these treaties, which originated in the mid-20th century, have remained largely static. Consequently, the principles codified in these agreements no longer fully reflect the realities of contemporary outer space activities.

A particularly pressing issue arises when considering the specific legal and regulatory challenges associated with space infrastructure projects, such as those envisioned in initiatives like HEIST or proposed by Taiwan. In these contexts, two critical aspects warrant particular attention: responsibility and liability. Responsibility pertains to the obligations of states and private entities under international law to ensure compliance with established norms and to avoid harmful activities in outer space. Liability addresses the mechanisms for attributing and compensating for damage caused by space objects, whether through collision, malfunction, or other forms of interference.

These challenges are compounded by the evolving nature of space activities, which often blur traditional lines of accountability and complicate the application of existing legal frameworks by creating a specific regime of responsibility in connection to private entities. The growing involvement of private entities, the increasing reliance on satellite infrastructure for critical services, and the rising threat of both cyber and kinetic attacks on space assets underscore the urgent need for updated legal instruments that can address the complexities of the New Space era. Without such updates, the foundational principles established by treaties like the OST risk becoming outdated and insufficient to govern the rapidly changing realities of outer space activities.

Responsibility and Attribution

Unlike the legal framework governing activities under the sea, where responsibility and liability are more unified, the framework for outer space distinguishes between the two concepts. This distinction stems from the fact that the OST provides only ‘principles’ rather than a comprehensive legal regime akin to the United Nations Convention on the Law of the Sea. Nevertheless, it would be incorrect to label the framework of the cosmic treaties (OST, Liability Convention,[51] and Registration Convention) ineffective simply because it is principle-based.

Article VI of the OST explicitly assigns responsibility to states for national activities in outer space, whether conducted by governmental or non-governmental entities. However, while this allocation of responsibility is relatively clear, the practical

[51] Heather S Fogo, ‘A Legal Mirage: State Responsibility for Non-State Actor Interference with Space Systems’ (2018) 55 Canadian Yearbook of International Law / Annuaire canadien de droit international 180, 195–205.

[52] ibid.

enforcement of these rules, particularly concerning non-state actors,[52] remains complex. Wang and Hu emphasize the necessity of distinguishing between the allocation of responsibility under Article VI of the OST and the attribution of acts under ARSIWA. According to these authors, the concept of allocation ensures that states maintain regulatory oversight over private entities, thereby fostering the development of space activities. Attribution, on the other hand, focuses on determining wrongdoing under international law. The difference is, according to Wang and Hu, significant, especially considering that ARSIWA was not developed at the time of the OST's creation.[53] Moreover, it can be argued that the space law regime may be considered *lex specialis* in relation to the general international law framework on state responsibility.

An interesting phenomenon arises from the fact that attribution under space law has a lower threshold than attribution under general international law. As Wang and Hu note, 'even if certain commercial space behaviours could be defined as "national activities" of state A under space law, it does not necessarily mean these acts could be attributed to state A under the law of state responsibility.'[54] General international law requires stricter standards for attribution to effectively use ARSIWA, which states that an internationally wrongful act must involve conduct attributable to a state and constitute a breach of an international obligation. For instance, under Articles 1, 2, 28, and 30 of ARSIWA, legal consequences include cessation, assurances of non-repetition, and reparations for wrongful acts. However, for private actors, these consequences hinge on whether their activities are attributable to a state.

Regarding attribution, it is important to highlight that Crawford argues that, in theory, every act of an individual could be attributable to a state due to the individual's connections to that state. Nevertheless, this broad approach is generally avoided in international law.[55] Instead, international law provides specific legal methods for attributing the actions of individuals to states. One such method is the overall control test, although its application in the context of outer space may present significant challenges.

This test allows for attribution without requiring specific directions for each activity, provided there is substantial evidence of state involvement, such as financial support, coordination, or operational assistance.[56] For instance, paying for commercial space services used by belligerents or coordinating joint operations could establish such

[53] G Wang and Y Hu, 'Allocation and Attribution of Commercial Space Activities in Armed Conflict' [Pre-Publication] (2025) 50 Air and Space Law 1–7. <https://kluwerlawonline.com/api/Product/CitationPDFURL?file=Journals\AILA\AILA2025001.pdf> accessed 11 January 2025.

[54] ibid 11.

[55] United Nations and James Crawford (eds), *The International Law Commission's Articles on State Responsibility: Introduction, Text, and Commentaries* (CUP 2002) 91.

[56] Antonio Cassese, 'The Nicaragua and Tadić Tests Revisited in Light of the ICJ Judgment on Genocide in Bosnia' (2007) 18 European Journal of International Law 649, 651.

involvement. Nonetheless, this approach contrasts with the stricter ‘effective control’ test, which could be seen as less applicable to the unique risks and dynamics of outer space and cybersecurity.

Critics of the lower threshold for attribution argue that it could unfairly result in an excessive number of commercial activities being deemed acts of the state, even without the state’s knowledge. However, this perspective overlooks the inherently hazardous nature of space activities, including their potential to trigger the Kessler Syndrome or other cascading risks.[57] As with nuclear regulation, space activities require high-risk oversight, as their consequences often transcend national boundaries. Granting states a blank check to disclaim responsibility would undermine global order.

Given the covert and unpredictable nature of space activities, particularly those with military implications, the ‘overall control’ test should be considered alongside ‘effective control’. This dual approach would urge states to rigorously supervise commercial activities under their jurisdiction without imposing unreasonable burdens. Such supervision could focus on entities that obtain operational licenses, ensuring compliance through regulatory measures embedded in national space policies.[58]

The differentiation between allocation and attribution is central to the legal architecture of space law. Allocation under the OST confirms which state bears international responsibility for a space activity, including obligations of authorization, supervision, and assurance, as well as any resulting legal consequences. By contrast, attribution under general international law determines whether a (non-governmental) entity’s space activity constitutes a state act, enabling the identification of ‘internationally wrongful acts’ and their associated legal consequences. Especially complex is the situation regarding private actors.

Wang and Hu illustrate this distinction through a hypothetical scenario: a satellite operated by Company M, licensed in State A, provides military services to State B in a conflict against State C. While State A is allocated responsibility under Article VI of the OST, State C would struggle to attribute the actions to State A under ARSIWA without meeting the higher threshold. Consequently, State C would be unable to pursue legal recourse or countermeasures against State A.[59]

The authors propose a standard that combines Article VI’s allocation principles with a state’s awareness of commercial activities. If there is a ‘generally close

[57] DJ Kessler and others, ‘The Kessler Syndrome: Implications to Future Space Operations’ (2010) <https://www.semanticscholar.org/paper/THE-KESSLER-SYNDROME%3A-IMPLICATIONS-TO-FUTURE-SPACE-Kessler-Johnson/227655e022441d1379dfdc395173ed2e776d54ee> accessed 12 January 2024; Mike Wall, ‘Kessler Syndrome and the Space Debris Problem’ (*Space.com*, 15 November 2021) <https://www.space.com/kessler-syndrome-space-debris> accessed 12 January 2024.

[58] Wang and Hu (n 53) 9–11.

[59] ibid 3–7.

connection' between the services and the state, and the state is aware of the activities, such involvement could be attributed to the state. This lower threshold encourages proactive supervision, preventing commercial space involvement from exacerbating international tensions and misperceptions.

This chapter argues that states must adopt sufficient regulatory frameworks regarding space's inherent risks and the dual challenges of responsibility and attribution. By doing so, they can ensure compliance with international obligations while fostering the sustainable development of outer space activities.

3. DISCUSSION – MALICIOUS ACTIVITIES ON THE INFRASTRUCTURES AND LEGAL CHALLENGES CONNECTED TO THEM

A. Non-Legal Considerations

When discussing malicious activities in space and under the sea, it is essential to address topics that extend beyond purely legal considerations, as they significantly impact law enforcement, *de lege lata* principles, and proper legal interpretation. For instance, subsea cables offer greater capacity and speed than space infrastructure. According to Frasca and Galantini, 'Were undersea cable networks to disappear, the entire capacity of the Earth's satellite network could handle just 7 per cent of the communications currently sent via cable from the United States alone.'[60] Consequently, a malicious act targeting satellites may cause less damage, while a well-executed attack on subsea cables could result in far-reaching consequences, necessitating different legal and policy responses.

In the aftermath of such attacks, repair processes also differ substantially between these two domains. Subsea cables, though located on the seabed, are relatively accessible for repair. Scuba divers or remotely operated vehicles can reach them, replace hardware, and restore functionality. By contrast, repairing space infrastructure presents much greater challenges. Repairs to satellites, whether in terms of hardware or orbital adjustments, are rare and often infeasible due to the inaccessibility of space. This highlights the greater vulnerability of space assets, as they cannot be easily reached or restored following an attack.

Examining the vectors of attack further illustrates these vulnerabilities. In space, attacks may include satellite-to-satellite engagements and ground-to-satellite disruptions such as jamming, spoofing, or eavesdropping. Furthermore, the cyberattack conducted by Turla in 2015 demonstrated that space assets could be exploited to conceal the location of command-and-control servers for malicious activities, thereby providing a shield

[60] F Cappelletti, A Nestoras and European Liberal Forum (eds), *Towards a New European Security Architecture* (European Liberal Forum 2023) 53.

of anonymity for threat actors.[61] Conversely, subsea cable attacks are more likely to involve physical damage to the cables or cyberattacks focusing on eavesdropping or espionage. These differing attack vectors necessitate tailored approaches to security and legal frameworks for each domain.

Moreover, it is essential to consider the dual-use nature of satellites as well as subsea cables. According to Azcárate Ortega, it is also necessary to distinguish between 'dual-use' and 'dual-purpose'. Dual-use refers to assets that can serve both military and civilian objectives, such as a global navigation satellite system (GNSS). By contrast, dual-purpose denotes cases where an asset is utilized for a purpose different from its originally intended function – for instance, a robotic arm designed for space debris removal being repurposed by the military to capture satellites instead of debris, or the weaponization of space debris.[62]

The dual-use nature of certain objects makes them potential military objectives and thus legitimate targets under the laws of armed conflict. However, the situation regarding subsea cables may be somewhat different, as the concept of dual-purpose applications is less readily applicable in this domain. The absence of a dual-use nature in subsea cables thus contributes to a clearer legal framework governing the undersea environment.

Another critical consideration is the attribution of cyberattacks, particularly when private or non-governmental entities are involved. Attribution remains a complex and debated topic among experts, including Vostoupal,[63] Wang,[64] and Li.[65] Due to the lack of consistent state practice, there is no universally accepted method for addressing this issue. While the 'effective control' test has been criticized for its overly strict threshold and is considered outdated, alternatives such as Li's 'virtual control test' or Margulies's proposal offer potential solutions. However, the feasibility and practical application of these tests remain open questions, underscoring the need for further academic and practical exploration of this issue.

B. Lack of Regulation Across Outer Space and Subsea Domains

One of the most pressing issues in the field of cybersecurity is the absence of adequate regulation, in both outer space and undersea contexts. Until recently, the

[61] 'Turla Hiding in the Sky: Russian Speaking Cyberespionage Group Exploits Satellites to Reach the Ultimate Level of Anonymity' (*Kaspersky*, 9 September 2015) <https://www.kaspersky.com/about/press-releases/turla-hiding-in-the-sky-russian-speaking-cyberespionage-group-exploits-satellites-to-reach-the-ultimate-level-of-anonymity> accessed 1 March 2025.

[62] AA Ortega, 'Not a Rose by Any Other Name: Dual-Use and Dual-Purpose Space Systems' [2023] (*Lawfare*, 5 June 2023) <https://www.lawfaremedia.org/article/not-a-rose-by-any-other-name-dual-use-and-dual-purpose-space-systems> accessed 1 March 2025.

[63] Vostoupal (n 11).

[64] Wang and Hu (n 53).

[65] Du Li, 'Legal Challenges of Attributing Malicious Cyber Activities against Space Activities' (2024) 37 Leiden Journal of International Law 963, 968.

regulation of cybersecurity in outer space remained largely unexplored, with experts calling for at least a basic framework to address this emerging challenge. However, this situation is gradually evolving, driven by bottom-up regulatory efforts, such as transnational private regulation and the adoption of soft law instruments. Despite these developments, clear definitions of these regulatory approaches remain elusive.

Soft law, as conceptualized by Dunk, refers to nonbinding realpolitik rules[66] adopted by traditional subjects of international law.[67] A notable example of soft law is the United Nations Resolution in the Context of International Security or ARSIWA,[68] which provides nonbinding yet influential guidance on state behaviour. On the other hand, transnational private regulation represents a system where international private entities take the lead in enforcing and shaping regulatory frameworks.[69] In the context of cybersecurity, standards[70] such as the NIST cybersecurity framework[71] exemplify this approach, offering guidance that is widely adopted by stakeholders across industries.

A parallel challenge exists under the sea, where subsea cables – critical infrastructure for global communications – are governed by a general legal framework but lack specific cybersecurity regulations.[72] Because of the plurality of legal regimes,[73] the legal analysis could be hard to navigate. The absence of targeted measures leaves these essential assets vulnerable to emerging threats, underscoring the need for specialized attention.[74]

In conclusion, cybersecurity regulation, both in outer space and for subsea cables, has long been insufficiently addressed in national strategies. Furthermore, state practice in these areas remains underdeveloped. Nevertheless, there is reason for cautious

[66] Frans von der Dunk, ‘Contradictio in Terminis or Realpolitik?’ in *Soft Law in Outer Space*, vol 102 (Böhlau Verlag 2012) <https://www.vr-elibrary.de/doi/10.7767/boehlau.9783205791850.31> accessed 27 July 2024.

[67] Vettorel (n 49).

[68] United Nations and James Crawford (eds), *The International Law Commission's Articles on State Responsibility: Introduction, Text, and Commentaries* (Cambridge University Press 2002).

[69] Fabrizio Cafaggi, ‘The Many Features of Transnational Private Rule-Making: Unexplored Relationships between Custom, Jura Mercatorum and Global Private Regulation’ (2015) 36 University of Pennsylvania Journal of International Law 875.

[70] NCCoE, ‘Cybersecurity for the Space Domain’ (*NIST*) <https://www.nccoe.nist.gov/cybersecurity-space-domain> accessed 1 February 2024; ‘P3349 - Space System Cybersecurity Working Group - The Project’ <https://sagroups.ieee.org/3349/the-project/> accessed 13 January 2024.

[71] NCCoE (n 70); M Scholl and T Suloway, ‘Introduction to Cybersecurity for Commercial Satellite Operations (2nd Draft)’ (National Institute of Standards and Technology 2022) NISTIR 8270 (Draft) <https://csrc.nist.gov/publications/detail/nistir/8270/draft> accessed 29 May 2022.

[72] Katherine Walla, ‘International Law Doesn't Adequately Protect Undersea Cables. That Must Change’ (*Atlantic Council*, 25 January 2024) <https://www.atlanticcouncil.org/content-series/hybrid-warfare-project/international-law-doesnt-adequately-protect-undersea-cables-that-must-change/> accessed 13 January 2025.

[73] Besch and Brown, ‘Securing Europe's Subsea Data Cables’ (n 6).

[74] Davenport (n 37) 82–83.

optimism, as the increasing focus on these issues may pave the way for more robust and comprehensive regulatory frameworks in the future, at least in the form of soft law or transnational private regulation. Unfortunately, under the current circumstances, a (binding) international treaty seems like a utopian dream.

4. CONCLUSION

Both systems under consideration – subsea cables and satellite infrastructure – feature significant complexity and unique vulnerabilities. However, complexity should not be viewed as an insurmountable obstacle. Instead, it underscores the importance of diversity and of maintaining a ‘Plan B’, as exemplified by initiatives like the HEIST project. Falco’s and Burnett’s[75] assertions that greater diversity enhances the security of data transfer and communication is particularly relevant in this context. Hybrid systems integrating both subsea cables and satellite networks can provide redundancy and resilience, mitigating risks associated with reliance on a single type of infrastructure.

In the context of outer space, the issue of attribution remains particularly intricate due to the allocation of responsibility, which plays a pivotal role in determining liability and accountability. If outer space is to serve as a viable alternative or complement to subsea cables, responsibility allocation mechanisms need to be substantially re-evaluated. Such a change could lead states to exercise caution in their engagement with space activities, particularly given that the standards for attribution under the law of state responsibility are stricter than those of space law.

Communication networks are classified as high-risk environments, necessitating rigorous security measures and international cooperation. Unlike traditional state-controlled activities, such as those associated with spaceports, the operation of subsea cables often involves private entities leasing infrastructure. This divergence introduces complex risk scenarios involving multiple stakeholders, including states and private corporations. Governance issues in these domains highlight the inadequacy of the current regulatory framework, particularly for submarine cables, which lack cybersecurity-specific measures.

To address these challenges, states and international organizations must adopt more harmonized legal frameworks. For subsea cables, this could involve updating UNCLOS to include specific cybersecurity provisions and improving enforcement mechanisms for existing obligations. For satellites, revising and modernizing the ‘cosmic treaties’ to reflect contemporary realities, including private sector involvement

[75] Burnett (n 46) 1665.

and cybersecurity threats, is crucial. These frameworks should promote proactive supervision and accountability while accommodating the unique risks of each domain.

Furthermore, initiatives like the HEIST project exemplify the potential for hybrid solutions that combine the strengths of both subsea and satellite infrastructure. These systems should be underpinned by robust legal and regulatory frameworks to ensure their effectiveness and resilience against emerging threats.

In conclusion, while the increased focus on these issues gives cause for cautious optimism, the road to achieving comprehensive international regulation remains challenging. Nonetheless, by leveraging hybrid systems, strengthening international cooperation, and updating legal frameworks, it is possible to enhance the resilience and security of global connectivity in the face of evolving cyber and geopolitical threats.

CyCon 2025: The Next Step
17th International Conference on Cyber Conflict
C. Kwan, N. Gratzer, K. Podiņš, M. Tolppa (Eds.)
2025 © NATO CCDCOE Publications, Tallinn

Permission to make digital or hard copies of this publication for internal use within NATO and for personal or educational use when for non-profit or non-commercial purposes is granted provided that copies bear this notice and a full citation on the first page. Any other reproduction or transmission requires prior written permission by NATO CCDCOE.

In or Out? Managing Risks from the UN Cybercrime Convention

Lisandra Novo
Law & Tech Counsel
Strategic Litigation Project, Atlantic Council
Washington, DC, United States

Abstract: The United Nations General Assembly unanimously adopted the first “global” treaty on cybercrime in December 2024. On the surface, it appears to be an international cooperation instrument for tackling cybercrime, but the treaty has been plagued with issues ever since Russia first proposed it. It poses grave risks to national security, global cybersecurity, and human rights alike. States faced a difficult choice in negotiations—to participate and try to shape the text in a consensus-based process, knowing they would have to compromise on matters of fundamental importance or refuse to join and watch as countries like Russia, Iran, and China manipulate the UN system.

Now, with the Convention opening for signature in 2025, and a negotiation process for a supplementary protocol on additional crimes envisioned in the text, States must choose once again whether to ratify it and participate in any future changes or abstain and watch as an already worryingly broad Convention inevitably grows and presents new threats. This paper will outline several of the risks the current text poses to national security, cybersecurity, and human rights through an analysis of specific provisions as well as insights from the corresponding negotiating history. In the second half, it will propose possible mitigation strategies to cope with these risks and examine their drawbacks. These strategies include refusing to ratify the treaty, urging cooperation under the Budapest Convention instead, and/or ratifying the Convention with reservations.

Keywords: *cybercrime, law of treaties, reservations, declarations, Budapest Convention*

1. INTRODUCTION

On December 24, 2024, United Nations officials touted the adoption of a "landmark convention," the "first international anti-crime treaty in 20 years," and a "major victory for multilateralism."[1] The UN General Assembly unanimously adopted the United Nations Convention Against Cybercrime; Strengthening International Cooperation for Combating Certain Crimes Committed by Means of Information and Communications Technology Systems and for the Sharing of Evidence in Electronic Form of Serious Crimes (UN Cybercrime Convention or the Convention).[2] The Convention's convoluted name is a much more accurate representation of how the international community should feel about this so-called achievement.

The negotiations were fraught with divisions among UN Member States. Industry experts and civil society organizations decried the text and the dangers that, if adopted, it would pose to national security, cybersecurity, and human rights. Significant compromises were made in the name of multilateralism, negotiations were delayed, and agreements were barely reached. In the end, the UN General Assembly chose to adopt the text as it was, presumably to not endanger the precarious balance that was achieved. Now that the Convention will open for signature, States must decide—are they in or out? In this paper, I outline some of the major concerns posed by the Convention on security and human rights, propose mitigation strategies that States should consider, and also list the drawbacks of these strategies. Regardless of the final choice, no State should jump into, or out of, this framework without a careful assessment, one likely to be as lengthy and painful as the Convention's name.

2. RISKS ALL AROUND

The UN Cybercrime Convention poses serious risks to national and cybersecurity as well as to human rights. Experts have warned that the Convention could allow States to force tech experts to assist in breaking encryption or to reveal source codes and that it is likely to enable human rights abuses. This section will cover a brief history of the Convention and present some of the most discussed risks. A detailed analysis of all possible risks is beyond the scope of this paper.

[1] Press Release, UN Office on Drugs and Crime, UN General Assembly Adopts Landmark Convention on Cybercrime (Dec. 24, 2024), https://www.unodc.org/unodc/en/press/releases/2024/December/un-general-assembly-adopts-landmark-convention-on-cybercrime.html. *See also* Vibhu Mishra, *UN General Assembly Adopts Milestone Cybercrime Treaty*, UN NEWS (Dec. 24, 2024), https://news.un.org/en/story/2024/12/1158521; *Making the Digital and Physical World Safer: Why the Convention Against Cybercrime Matters*, UN NEWS (Dec. 24, 2024), https://news.un.org/en/story/2024/12/1158526.

[2] U.N. General Assembly, Resolution adopted by the General Assembly on 24 December 2024, U.N. Doc. A/RES/79/243 (Dec. 31, 2024) [hereinafter UN Cybercrime Convention].

A. Where Did It Come From?

The framing employed by UN officials suggests the Convention emerged from a barren landscape, crafted by the sheer strength of multilateralism. But previous international treaties on criminal matters from 20 years ago also deal with international cooperation. The UN Convention Against Transnational Organized Crime was adopted by the UN General Assembly in 2000,[3] and the UN Convention Against Corruption was adopted in 2003.[4] These served as important precedents in negotiations for the UN Cybercrime Convention.[5]

But the real basis, and trigger, for the UN Cybercrime Convention is not a UN treaty at all—it is the Council of Europe (CoE) Convention on Cybercrime (the Budapest Convention).[6] Many of its provisions were reproduced in the draft UN Cybercrime Convention.[7] The CoE (not to be confused with the EU) is a European intergovernmental organization founded in 1949[8] with the mission to "promote democracy, human rights and the rule of law across Europe and beyond."[9] Russia was a member, until it left hours before a vote on whether to expel it because of its full-scale invasion of Ukraine.[10]

Russia refused to join the Budapest Convention, alleging its framework would violate its sovereignty and the principle of non-intervention,[11] and criticized the CoE for limiting participation to a "select club of 'developed democracies', the door to

[3] *UN Convention Against Transnational Organized Crime and the Protocols Thereto*, UN OFFICE DRUGS CRIME, https://www.unodc.org/unodc/en/organized-crime/intro/UNTOC.html (last visited Jan. 3, 2025).

[4] *UNCAC*, UN OFFICE DRUGS CRIME, https://www.unodc.org/corruption/en/uncac/index.html (last visited Jan. 3, 2025).

[5] *See, e.g.*, *(11th Meeting) Reconvened Concluding Session of the Ad Hoc Committee to Elaborate a Comprehensive International Convention on Countering the Use of Information and Communications Technologies for Criminal Purposes*, UN WEB TV (Aug. 5, 2024), https://webtv.un.org/en/asset/k1u/k1ug11vydh.

[6] Convention on Cybercrime, Nov. 23, 2001, T.I.A.S. 13174, E.T.S. No. 185 [hereinafter Budapest Convention].

[7] *See* Briefing Note, Conventions on Cybercrime: The Budapest Convention and the Draft UN Treaty, Council of Europe, Cybercrime Division (Aug. 27, 2024), https://rm.coe.int/conventions-on-cybercrime-the-budapest-convention-and-the-draft-un-tre/1680b1631a; *United Nations Treaty on Cybercrime Agreed by the Ad Hoc Committee*, COUNCIL EUR. (Aug. 8, 2024), https://www.coe.int/en/web/cybercrime/-/united-nations-treaty-on-cybercrime-agreed-by-the-ad-hoc-committee.

[8] *The Council of Europe and the European Union*, COUNCIL EUR., https://www.coe.int/en/web/portal/european-union (last visited Jan. 4, 2025); *46 Member States*, COUNCIL EUR., https://www.coe.int/en/web/portal/46-members-states (last visited Jan. 4, 2025).

[9] *The Council of Europe at a Glance*, COUNCIL EUR., https://www.coe.int/en/web/portal/the-council-of-europe-at-a-glance (last visited Jan. 4, 2025).

[10] *Russia Quits Council of Europe Rights Watchdog*, REUTERS (Mar. 15, 2022), https://www.reuters.com/world/europe/russia-formally-quits-council-europe-rights-watchdog-2022-03-15/.

[11] Karine Bannelier & Eugenia Lostri, *Is Anyone Happy With the UN Cybercrime Convention?*, LAWFARE (Dec. 2, 2024), https://www.lawfaremedia.org/article/is-anyone-happy-with-the-un-cybercrime-convention; Joyce Hakmeh, *A New UN Cybercrime Treaty? The Way Forward for Supporters of an Open, Free, and Secure Internet*, CFR (Jan. 13, 2020), https://www.cfr.org/blog/new-un-cybercrime-treaty-way-forward-supporters-open-free-and-secure-internet.

which is cracked only by invitation."[12] It was Russia that proposed the creation of a UN Cybercrime Convention in the first place.[13] In 2017, it presented a letter to the UN General Assembly, attaching its own draft convention, and then, in 2019, the resolution it sponsored, with Belarus, China, the Islamic Republic of Iran (IRI), Nicaragua, and others, passed in the UN General Assembly, paving the way for the Ad Hoc Committee that would engage in years of negotiation and deliver the final text of the Convention in August 2024.[14] It will open for signature in October 2025 and requires 40 ratifications to enter into force.[15] However long that process takes, States must immediately begin to assess whether to ratify or not.

B. Broad Scope and Reach

The lengthy title of the Convention is the first sign of trouble. There was a clear divide on the vision for the Convention,[16] which Russia characterized as "[t]he civilizational clash between the neoliberal collective West and its satellites and a large part of the world majority."[17] Ultimately, a broad vision prevailed, one the EU representative noted was unprecedented in scope both for the inclusion of "serious crimes" and the real-time data collection and interception measures.[18]

Article 2(h) of the Convention defines a "serious crime" as "conduct constituting an offence punishable by a maximum deprivation of liberty of at least four years or a more serious penalty."[19] The inclusion of a category of crimes not established by the Convention and merely defined by reference to domestic law makes any other problematic provision exponentially more so, given the increased potential for abuse. For example, in Iran, charges such as "corruption on Earth" and "propaganda against the State" carry the death penalty, fulfilling the serious crime requirement, and have

[12] Pyotr Litvishko, *The First Global Treaty against Cybercrime: From Geopolitical Confrontation Towards Professional Compromise*, EMBASSY RUSSIAN FEDERATION IN REPUBLIC SOUTH AFRICA, https://russianembassyza.mid.ru/en/press-centre/news/the_first_global_treaty_against_cybercrime_from_geopolitical_confrontation_towards_professional_comp/ (last visited Jan. 4, 2025).

[13] Andrew C. Adams & Daniel Podair, *Confusion & Contradiction in the UN "Cybercrime" Convention*, LAWFARE (Dec. 9, 2024), https://www.lawfaremedia.org/article/confusion---contradiction-in-the-un--cybercrime--convention.

[14] *Ad Hoc Committee to Elaborate a Comprehensive International Convention on Countering the Use of Information and Communications Technologies for Criminal Purposes*, UN OFFICE DRUGS CRIME, https://www.unodc.org/unodc/en/cybercrime/ad_hoc_committee/home (last visited Jan. 3, 2025); Karen Gullo & Katitza Rodriguez, UN Cybercrime Treaty Timeline, EFF, https://www.eff.org/pages/un-cybercrime-treaty-timeline (last visited Dec. 25, 2024).

[15] *United Nations Convention Against Cybercrime; Strengthening International Cooperation for Combating Certain Crimes Committed by Means of Information and Communications Technology Systems and for the Sharing of Evidence in Electronic Form of Serious Crimes*, UN OFFICE DRUGS CRIME, https://www.unodc.org/unodc/en/cybercrime/convention/home.html (last visited Jan. 4, 2025).

[16] *See, e.g.*, *(4th Meeting) Reconvened Concluding Session of the Ad Hoc Committee to Elaborate a Comprehensive International Convention on Countering the Use of Information and Communications Technologies for Criminal Purposes*, UN WEB TV (July 30, 2024), https://webtv.un.org/en/asset/k1q/k1q07knzha.

[17] Litvishko, *supra* note 12.

[18] *(12th Meeting) Reconvened Concluding Session of the Ad Hoc Committee to Elaborate a Comprehensive International Convention on Countering the Use of Information and Communications Technologies for Criminal Purposes*, UN WEB TV (Aug. 5, 2024), https://webtv.un.org/en/asset/k16/k1664l3756.

[19] UN Cybercrime Convention, *supra* note 2, art. 2(h).

been used to prosecute people for online activity.[20] The UK criticized this definition, which is entirely dependent on domestic law, for making it "impossible" to predict what would constitute serious crimes.[21] In Rwanda's view, the baseline of a minimum penalty of four years of incarceration does not constitute a serious crime at all.[22]

There are also significant risks related to the offenses that the Convention requires each State Party to criminalize.[23] Specifically, the intent language for several provisions means that good-faith security researchers could be prosecuted for accessing an "[ICT] system without right."[24] Katitza Rodriguez of the Electronic Frontier Foundation underscored that "[r]esearchers around the world are routinely threatened or prosecuted for merely exposing security breaches in order to get them fixed."[25] Experts also warned that provisions allowing States Parties to force IT employees, or seemingly anyone with special knowledge about a system, to provide information to access that system or to seize or secure data could threaten cyber and data security globally.[26] The International Chamber of Commerce cautioned that this provision "could even be interpreted to include compelled disclosure of previously unknown vulnerabilities, private encryption keys, or proprietary information like source code[s]."[27]

C. Safeguards Susceptible to Abuse

Many States valiantly fought to protect the human rights safeguards from further degradation during the final negotiations.[28] However, the result was as much a compromise as the rest of the text. Article 6, "Respect for human rights," is a general provision that applies to the whole text.[29] Article 24, "Conditions and safeguards," is included under Chapter IV, "Procedural measures and law enforcement," and thus applies only to those provisions.[30] The safeguards to be adopted under this provision are those provided under a Party's domestic law and in accordance with its obligations under international human rights law.[31] The Convention thus does not impose its own

[20] *See, e.g.*, Press Release, UN OHCHR, Iran: UN Experts Alarmed by Death Sentence Imposed on Peaceful Activist, Demand Moratorium on Death Penalty (May 13, 2024), https://www.ohchr.org/en/press-releases/2024/05/iran-un-experts-alarmed-death-sentence-imposed-peaceful-activist-demand.

[21] *(12th Meeting) Reconvened Concluding Session of the Ad Hoc Committee* (Aug. 5, 2024), *supra* note 18.

[22] *(4th Meeting) Reconvened Concluding Session of the Ad Hoc Committee* (July 30, 2024), *supra* note 16.

[23] UN Cybercrime Convention, *supra* note 2, arts. 7–21.

[24] *Id*. art. 7(1).

[25] Katitza Rodriguez, *The UN Cybercrime Convention: Analyzing the Risks to Human Rights and Global Privacy*, JUST SECURITY (Aug. 27, 2024), https://www.justsecurity.org/98738/cybercrime-convention-human-rights/.

[26] *See id*.; Nick Ashton-Hart, *A New U.N. Cybercrime Convention Could Land You in Jail*, DC JOURNAL (Aug. 26, 2024), https://dcjournal.com/a-new-u-n-cybercrime-convention-could-land-you-in-jail/.

[27] *(4th Meeting) Reconvened Concluding Session of the Ad Hoc Committee* (July 30, 2024), *supra* note 16 (oral statement of the International Chamber of Commerce); *see also id*. (oral statement of Microsoft).

[28] *See, e.g.*, *(11th Meeting) Reconvened Concluding Session of the Ad Hoc Committee* (Aug. 5, 2024), *supra* note 5.

[29] UN Cybercrime Convention, *supra* note 2, art. 6.

[30] *Id*. art. 24; Rodriguez, *supra* note 25.

[31] UN Cybercrime Convention, *supra* note 2, art. 24(1–2).

minimum human rights safeguards. While this opens the door to abuse, it also allows States to include safeguards that afford greater protection.

The US noted the Convention allows for the refusal of mutual legal assistance requests "that discriminate on the basis of sex, race, language, religion, nationality, ethnic origin, or political opinion" and that it "will reject such requests and will do everything in [its] power to ensure that others do as well."[32] Rejection is not so simple. Article 40 has been criticized for its optional nature and high threshold before a request can be rejected.[33] Rather than imposing an obligation to reject such requests, the Convention simply mentions that a State is not prevented from doing so, but it must have "substantial grounds" to believe a request was made to prosecute or punish someone on a protected ground.[34] The same structure, and criticism applies to the refusal of extradition requests.[35] Negotiators tried to include an exception for political offenses, meaning States would not be required to comply with extradition requests for political crimes, a common inclusion in extradition and mutual legal assistance treaties (MLATs),[36] but strong opposition ultimately prevailed.[37]

Certain States were deeply opposed to safeguards of any form. On the day the Convention was finally adopted by the Ad Hoc Committee, the Iranian delegation initiated a vote to remove human rights provisions, references to gender, and essential safeguards.[38] This is the same government a UN Fact-Finding Mission found responsible for crimes against humanity, including gender persecution, in the context of the Woman, Life, Freedom protests sparked by Mahsa Jina Amini's death

[32] *Explanation of Position of the United States on the Adoption of the Resolution on the UN Convention Against Cybercrime in the UN General Assembly's Third Committee*, US MISSION TO UN (Nov. 11, 2024), https://usun.usmission.gov/explanation-of-position-of-the-united-states-on-the-adoption-of-the-resolution-on-the-un-convention-against-cybercrime-in-ungas-third-committee/.

[33] *See, e.g.*, Rodriguez, *supra* note 25.

[34] UN Cybercrime Convention, *supra* note 2, art. 40(22).

[35] *Id*. art. 37(15).

[36] *See (4th Meeting) Reconvened Concluding Session of the Ad Hoc Committee* (July 30, 2024), *supra* note 16. *See also* Extradition Treaty between the US and the Philippines (1995), https://www.congress.gov/104/cdoc/tdoc16/CDOC-104tdoc16.pdf; Mutual Legal Assistance, Extradition and Recovery of Proceeds of Corruption in Asia and the Pacific: Frameworks and Practices in 27 Asian and Pacific Jurisdictions, ADB/OECD Anti-Corruption Initiative for Asia and the Pacific (2008), https://www.oecd.org/content/dam/oecd/en/publications/reports/2008/03/mutual-legal-assistance-extradition-and-recovery-of-proceeds-of-corruption-in-asia-and-the-pacific-frameworks-and-practices-in-27-asian-and-pacific-jurisdictions-final-report_g1gha805/9789264043701-en.pdf.

[37] *See, e.g.*, Ad Hoc Committee to Elaborate a Comprehensive International Convention on Countering the Use of Information and Communications Technologies for Criminal Purposes, Compilation of Views Submitted by Member States on the Scope, Objectives and Structure (Elements) of a Comprehensive International Convention on Countering the Use of Information and Communications Technologies for Criminal Purposes, Note by the Secretariat (Nov. 17, 2021), 28 (views of Egypt), 61 (views of the UK), U.N. Doc. A/AC.291/4; Rodriguez, supra note 25.

[38] *(16th Meeting) Reconvened Concluding Session of the Ad Hoc Committee to Elaborate a Comprehensive International Convention on Countering the Use of Information and Communications Technologies for Criminal Purposes*, UN WEB TV (Aug. 8, 2024), https://webtv.un.org/en/asset/k12/k1207pa2nz?kalturaStartTime=6095&kalturaStartTime=5514; Lisandra Novo, *The UN Finally Advances a Convention on Cybercrime ... and No One Is Happy About It*, NEW ATLANTICIST (Aug. 14, 2024), https://www.atlanticcouncil.org/blogs/new-atlanticist/the-un-finally-adopts-a-convention-on-cybercrime-and-no-one-is-happy/.

in custody.[39] The attempt to remove references to gender in a treaty covering non-consensual dissemination of intimate images,[40] a crime that disproportionately affects women,[41] rings even more hollow.

The risk of using the Convention for transnational repression (TNR) has also been raised.[42] TNR is the practice of one State intimidating, harassing, threatening, or harming critics and dissidents residing in another country or their families back in their country of origin.[43] Sometimes subsumed under foreign interference,[44] this is a distinct practice targeting individuals and can include violent acts against them or their families.[45] Interpol notices are notorious for this kind of abuse[46] and receive additional scrutiny from some States like Canada.[47] Under the UN Cybercrime Convention, IRI officials, for example, could send a request to the country a dissident lives in for real-time traffic or content data,[48] instead of relying on threat actors to mount elaborate phishing campaigns to target the person.[49]

3. MITIGATION

Despite the Convention's problems, there is no doubt that cybercrime must be addressed. It poses grave national security risks[50] and has devastating consequences

[39] Press Release, UN OHCHR, Iran: Institutional Discrimination Against Women and Girls Enabled Human Rights Violations and Crimes Against Humanity in the Context of Recent Protests, UN Fact-Finding Mission Says (Mar. 8, 2024), https://www.ohchr.org/en/press-releases/2024/03/iran-institutional-discrimination-against-women-and-girls-enabled-human.

[40] UN Cybercrime Convention, *supra* note 2, art. 16

[41] Pavlina Pavlova, *Gendered Harms of Data Weaponization: Historical Patterns, New Battlefields, and the Implications for Democracy and National Security*, NEW AMERICA (Nov. 14, 2024), https://www.newamerica.org/future-security/reports/gendered-harms-of-data-weaponization/.

[42] *See, e.g.*, Deborah Brown, *New UN Cybercrime Treaty Primed for Abuse: States Should Reject Ratifying Convention on Human Rights Grounds*, HUMAN RIGHTS WATCH (Dec. 30, 2024), https://www.hrw.org/news/2024/12/30/new-un-cybercrime-treaty-primed-abuse.

[43] *See, e.g.*, *Transnational Repression*, FBI, https://www.fbi.gov/investigate/counterintelligence/transnational-repression (last visited Jan. 6, 2025).

[44] *See, e.g.*, Marie-Josée Hogue, Public Inquiry into Foreign Interference in Federal Electoral Processes and Democratic Institutions—Initial Report (May 3, 2024), 83, https://foreigninterferencecommission.ca/fileadmin/user_upload/Foreign_Interference_Commission_-_Initial_Report__May_2024__-_Digital.pdf; *Special Report 2022, Canada: Transnational Repression Host Country Case Study*, FREEDOM HOUSE (2022), https://freedomhouse.org/report/transnational-repression/canada.

[45] *Transnational Repression*, FREEDOM HOUSE, https://freedomhouse.org/report/transnational-repression (last visited Jan. 6, 2025).

[46] *See, e.g.*, Katitza Rodriguez, *The UN Cybercrime Draft Convention Remains Too Flawed to Adopt*, EFF (June 7, 2024), https://www.eff.org/am/deeplinks/2024/06/un-cybercrime-draft-convention-remains-too-flawed-adopt; Kate Robertson, *The UN's New Cybercrime Treaty Is Poised to Become a Vehicle for Complicity in the Global Mercenary Spy Trade*, LAWFARE (Aug. 22, 2024), https://www.lawfaremedia.org/article/a-global-treaty-to-fight-cybercrime-without-combating-mercenary-spyware.

[47] *Special Report 2022, Canada: Transnational Repression Host Country Case Study, supra* note 44.

[48] UN Cybercrime Convention, *supra* note 2, arts. 29(1), 30(1).

[49] *See, e.g.*, *Iran*, THE CITIZEN LAB (Dec. 11, 2024), https://citizenlab.ca/case-studies/iran/.

[50] *See, e.g.*, Simon Handler & Liv Rowley, *The 5×5—Cybercrime and National Security*, THE 5X5 (June 29, 2022), https://www.atlanticcouncil.org/commentary/the-5x5-cybercrime-and-national-security/.

for a wide array of victims.[51] States must now decide whether the UN is the best framework through which to combat cybercrime and opt in or, if the risks outweigh the benefits, opt out.

A. Opt Out

Civil society organizations and industry experts alike have strongly urged States to opt out.[52] They have expressed the concerns outlined in Section 2 for years[53] and maintain that no treaty is better than a "bad" one.[54] For some States, this may indeed be the case. Take the US, for example, where some of the world's biggest tech companies are headquartered and hold evidence ranging from communications between perpetrators to video or photo evidence capturing the commission of crimes.[55] While MLATs are the traditional channels for exchanging information, the US enacted the Clarifying Lawful Overseas Use of Data (CLOUD) Act in March 2018, in part because of the dramatic increase in requests for electronic evidence and the need to speed up response time.[56]

The number of requests would inevitably increase even more dramatically if the US joined the UN Cybercrime Convention. Further, the CLOUD Act is designed to assist countries "that have robust protections for privacy and civil liberties."[57] That reasoning would not apply to some of the potential parties to the UN Cybercrime Convention. Perhaps for these reasons, and because of the vocal opposition from industry and human rights groups, the US has indicated it is "unlikely to sign or ratify [the Convention] unless and until [they] see implementation of meaningful human rights and other legal protections by the convention's signatories."[58]

[51] *See, e.g.*, Internet Organised Crime Threat Assessment (IOCTA) 2024, Europol (2024), https://www.europol.europa.eu/cms/sites/default/files/documents/Internet%20Organised%20Crime%20Threat%20Assessment%20IOCTA%202024.pdf; Emily Ferguson & Emma Schroeder, This Job Post Will Get You Kidnapped: A Deadly Cycle of Crime, Cyberscams, and Civil War in Myanmar, Issue Brief (Nov. 2023), Cyber Statecraft Initiative, DFRLab, https://dfrlab.org/2023/11/13/this-job-post-will-get-you-kidnapped/.

[52] *See, e.g.*, *Global Business Urges Governments to Reject New International Cybercrime Treaty*, INT'L CHAMBER COMMERCE (Aug. 13, 2024), https://iccwbo.org/news-publications/news/global-business-urges-governments-to-reject-new-international-cybercrime-treaty/; Deborah Brown, *New UN Cybercrime Treaty Primed for Abuse, supra* note 42.

[53] *See, e.g.*, Katitza Rodriguez & George Wong, *Letter to the United Nations to Include Human Rights Safeguards in Proposed Cybercrime Treaty*, EFF (Feb. 27, 2022), https://www.eff.org/deeplinks/2022/02/letter-united-nations-include-human-rights-safeguards-proposed-cybercrime-treaty; Christian Ohanian, *The UN Cybercrime Treaty Has a Cybersecurity Problem in It*, JUST SECURITY (Oct. 17, 2022), https://www.justsecurity.org/83582/the-un-cybercrime-treaty-has-a-cybersecurity-problem-in-it/.

[54] *See, e.g.*, *UN Cybercrime Convention: FAQ on Necessary Reforms*, ACCESS NOW (Jan. 25, 2024), https://www.accessnow.org/guide/faq-un-cybercrime-convention-ahc/; Jonathan Greig, *Controversial UN Cybercrime Treaty Clears Final Hurdle Before Full Vote as US Defends Support*, RECORDED FUTURE NEWS (Nov. 12, 2024), https://therecord.media/un-cybercrime-treaty-clears-vote.

[55] *See, e.g.*, Michael A. Becker, The Gambia v. Facebook: Obtaining Evidence for Use at the International Court of Justice (Part I), EJIL: TALK! (Oct. 5, 2021), https://www.ejiltalk.org/the-gambia-v-facebook-obtaining-evidence-for-use-at-the-international-court-of-justice-part-i/.

[56] *CLOUD Act Resources*, U.S. DEP'T JUSTICE, https://www.justice.gov/criminal/cloud-act-resources (last updated Oct. 24, 2023).

[57] *Id*.

[58] *Explanation of Position of the United States on the Adoption of the Resolution on the UN Convention Against Cybercrime, supra* note 32.

1) Other Frameworks

A State that has decided not to ratify the UN Cybercrime Convention still has options. It can continue to use MLATs, but those only apply between signatories and are usually bilateral instruments. As discussed in Section 2.A, however, there is another global cybercrime treaty available: the Budapest Convention. Many countries have modeled their domestic cybercrime legislation on it,[59] and most European countries have ratified it, as have the US and Canada.[60] Many non-Western countries have joined or sought to do so since the start of the UN Cybercrime Convention negotiations. For example, in 2023 and 2024, nine countries ratified the Budapest Convention,[61] and 15 have requested to accede since 2020.[62] However, despite references to the Budapest Convention as the "gold standard," it too has been criticized on human rights grounds.[63]

2) Risks of Sitting Out

Russia, China, and Iran, among others, are not going to join the Budapest Convention.[64] For States that maintain close ties with them, the Budapest Convention does not solve the legal cooperation issue. Further, even if States see no need to cooperate with Russia, China, or the IRI, there are still many other UN Member States that are not party to the Budapest Convention. Of the current 78 State Parties to the Budapest Convention, 45 are CoE members (European), while 33 are not (non-European).[65] Of the 193 UN Member States,[66] 115 are not party to the Budapest Convention. Even accounting for countries that are signatories or are in the accession process (17),[67] 98 UN Member States remain outside this framework. If many countries do not ratify the UN Cybercrime Convention, a similar split could be replicated there.

The biggest risk of opting out is possibly watching authoritarian States mold the UN Cybercrime Convention to their vision of cyber governance and reintroduce

[59] *Cybercrime: Achievements*, COUNCIL EUR., https://www.coe.int/en/web/cybercrime/achievements (last visited Jan. 3, 2025); Dominik Zachar, *Battling Cybercrime Through the New Additional Protocol to the Budapest Convention*, NATO CCDCOE (2021), https://ccdcoe.org/library/publications/battling-cybercrime-through-the-new-additional-protocol-to-the-budapest-convention/.

[60] Treaty Office, *Chart of Signatures and Ratifications of Treaty 185*, COUNCIL EUR., https://www.coe.int/en/web/conventions/full-list?module=signatures-by-treaty&treatynum=185 (last visited Feb. 26, 2025) [hereinafter Budapest Convention, States Parties and Reservations].

[61] *Id*.

[62] Treaty Office, *Non-Member States of the Council of Europe: Five Years Validity of an Invitation to Sign and Ratify or to Accede to the Council of Europe's Treaties*, COUNCIL EUR. (Dec. 17, 2024), https://rm.coe.int/16806cac22.

[63] Deborah Brown, *Cybercrime Is Dangerous, But a New UN Treaty Could Be Worse for Rights*, JUST SECURITY (Aug. 13, 2021), https://www.justsecurity.org/77756/cybercrime-is-dangerous-but-a-new-un-treaty-could-be-worse-for-rights/; Issue Paper, Commissioner for Human Rights, Council of Europe, The Rule of Law on the Internet and in the Wider Digital World (2014), 16–17, 22–23, 93–107 https://rm.coe.int/16806da51c.

[64] Hakmeh, *supra* note 11; Adams & Podair, *supra* note 13.

[65] Budapest Convention, States Parties and Reservations, *supra* note 60.

[66] *Growth in United Nations Membership*, UNITED NATIONS, https://www.un.org/en/about-us/growth-in-un-membership (last visited Jan. 4, 2025).

[67] Budapest Convention, States Parties and Reservations, *supra* note 60.

provisions that had previously been eliminated.[68] The US observed the Convention "requires critical safeguards for the use of domestic powers like search and seizure or interception, including when providing mutual legal assistance," and warned that a "Party that does not provide for such safeguards … or whose safeguards are not in accordance with its international human rights law obligations," would be in breach of its obligations and urged all countries to reject such non-compliant requests.[69] But if rights-respecting States joining the Convention are in the minority, who will monitor compliance or raise it at the Conference of States Parties? And if the makeup of the Conference is tilted toward one group, these States may reach the 60-State threshold needed to begin negotiations on the proposed additional protocol that was the subject of contentious debate, which could include additional offenses, such as extremism and terrorism that are often vague and ill-defined.[70]

Russia has already raised two important additional risks of non-participation: collecting evidence on behalf of other countries and setting international precedents. Regarding the first, a Russian official warns of the Convention's "possible bad-faith instrumentalization for political and military purposes" by Ukraine and allies by collecting electronic evidence "against the Russian Federation," which would then be "indirectly, exfiltrated via various proxies and under the guise of unrelated proceedings on ordinary-law crimes."[71] While this example predictably posits Ukraine as the culprit, any country could make use of this same strategy to obtain data intended to abuse dissidents, critics, and others. On the latter, the same official opined that this Convention "surpassed" the Budapest Convention with respect to the interpretation of "investigation," because, unlike the Budapest Convention, the UN text allegedly encompasses "both investigative actions and proactive covert operational search measures," that is, "at the stages of detection, prevention and frustration of criminal offences."[72] A norm that would expand investigatory and prosecutorial powers to the stage before a crime is even committed requires extensive debate from the international community.

B. Opt In

Any country wishing to ratify the UN Cybercrime Convention will face both capacity issues and the requirement to implement changes to their domestic systems. Countries that are in the process of acceding to the Budapest Convention may face twice the

[68] *See, e.g.*, *A Discussion on the UN Cybercrime Convention*, CSIS (Oct. 4, 2024), https://www.csis.org/events/discussion-un-cybercrime-convention; Anja P. Jakobi & Lena Herbst, *Between a Rock and a Hard Place: The UN Cybercrime Convention*, PRIF SPOTLIGHT (Dec. 9, 2024), https://blog.prif.org/2024/12/09/between-a-rock-and-a-hard-place-the-un-cybercrime-convention/. *See also* Arun Sukumar & Arindrajit Basu, *Back to the Territorial State: China and Russia's Use of UN Cybercrime Negotiations to Challenge the Liberal Cyber Order*, J. CYBER POL'Y 1 (Dec. 13, 2024).

[69] *Explanation of Position of the United States on the Adoption of the Resolution on the UN Convention Against Cybercrime, supra* note 32.

[70] UN Cybercrime Convention, *supra* note 2, arts. 57(5)(g), 62(1); Rodriguez, *supra* note 25. *See (4th Meeting) Reconvened Concluding Session of the Ad Hoc Committee* (July 30, 2024), *supra* note 16.

[71] Litvishko, *supra* note 12.

[72] *Id*.

burden if they decide to join both. It is imperative to have a consistent approach to both frameworks. Given the significant risks posed by the UN Cybercrime Convention, any country considering ratification must assess its domestic system carefully and take advantage of the opportunities to define its own obligations. In short, States should lodge reservations and ensure domestic laws match their scope. The reservations to the Budapest Convention offer useful examples.

1) Budapest Convention and Its Many Reservations

Reservations are common practice in international treaties—as expressly addressed under the Vienna Convention on the Law of Treaties (VCLT).[73] The International Law Commission defines a reservation as a "unilateral statement, however phrased or named" that "purports to exclude or to modify the legal effect of certain provisions of the treaty in their application."[74] Declarations are statements that only purport to set out a State's understanding of a provision's scope.[75] Because the recommendation set out in this paper is to modify the legal effect of provisions, I will refer only to reservations.

Reservations are often seen as a way for a State to excuse itself from treaty obligations, such as the jurisdiction of the International Court of Justice in the settlement of disputes.[76] Reservations are understandably seen in a negative light in human rights treaties and are impermissible if they are against the treaty's object and purpose.[77] However, reservations can bind a State to a legal standard that is more protective than that included in the treaty. There has not been much study of this protective effect, but it is imperative for mitigating the risks posed by the UN Cybercrime Convention. A study of reservations to the Budapest Convention is illustrative. Table I shows that 42 (54%) of the Budapest Convention States Parties have lodged either a reservation, a declaration, or both, while 36 (46%) have not.[78] Of those 42, 19 (45%) lodged both reservations and declarations, 15 (36%) lodged only reservations, and eight (19%) lodged only declarations.

[73] Vienna Convention on the Law of Treaties, May 23, 1969, arts. 19–23, 1155 U.N.T.S. 331 (1969) [hereinafter VCLT].

[74] Guide to Practice on Reservations to Treaties, International Law Commission, 63d Sess., U.N. Doc. A/66/10 (2011), at 19, 1.1 (definition of reservations); *compare with* VCLT, *supra* note 73, art. 2(d).

[75] *Id*. at 21, 1.2 (definition of interpretative declarations).

[76] *See, e.g.*, Allegations of Genocide Under the Convention on the Prevention and Punishment of the Crime of Genocide (Ukraine v. Russian Federation), Order, Admissibility of the Declarations of Intervention, 2023 I.C.J. Rep. 354, 93–98 (June 5).

[77] *See, e.g.*, Reservations to the Convention on Genocide, Advisory Opinion, 1951 I.C.J. Rep. 15 (May 28); Boyes *et al., Social Pressure in the International Human Rights Regime: Why States Withdraw Treaty Reservations*, 54 Brit. J. Pol. Sci. 241 (2024).

[78] Data compiled by author from Budapest Convention, States Parties and Reservations, *supra* note 60.

TABLE I: RESERVATIONS AND DECLARATIONS BY STATES PARTIES TO THE BUDAPEST CONVENTION

Lodged Both	Only Reservations	Only Declarations	None
Andorra	Argentina	Brazil	Albania
Azerbaijan	Australia	Costa Rica	Armenia
Belgium	Austria	Georgia	Benin
Canada	Bulgaria	Iceland	Bosnia and Herzegovina
Chile	Colombia	Netherlands	Cabo Verde
Czech Republic	Greece	Portugal	Cameroon
Denmark	Israel	Republic of Moldova	Côte d'Ivoire
Finland	Latvia	Spain	Croatia
France	Montenegro		Cyprus
Germany	Nigeria		Dominican Republic
Hungary	Norway		Ecuador
Japan	Poland		Estonia
Liechtenstein	Sri Lanka		Fiji
Lithuania	Sweden		Ghana
Peru	United Kingdom		Grenada
Slovak Republic			Italy
Switzerland			Kiribati
Ukraine			Luxembourg
United States of America			Malta
			Mauritius
			Monaco
			Morocco
			North Macedonia
			Panama

Lodged Both	Only Reservations	Only Declarations	None
			Paraguay
			Philippines
			Romania
			Rwanda
			San Marino
			Senegal
			Serbia
			Sierra Leone
			Slovenia
			Tonga
			Tunisia
			Türkiye
19	**15**	**8**	**36**

The recommendations in this paper are also addressed to any country considering acceding to the Budapest Convention or that has already done so without reservations. A breakdown of the countries that lodged statements by geographical positioning shows that they are mostly European and high-income countries. Of the total 78 State Parties to the Budapest Convention, 45 (58%) are CoE members.[79] Of these 45, 29 (64%) lodged a reservation. On the other hand, of the 33 (42%) States Parties who are not CoE members (non-European countries), only 13 (39%) lodged a reservation, while the majority, 20 (61%), did not.[80] Here we see a completely inverse trend. The five countries that lodged the most reservations and declarations overall, as seen in Table II, were the US, Japan, Switzerland, Chile, and Peru.[81] Meanwhile, Nigeria is the only African State Party to have lodged either a reservation or declaration.[82]

[79] Budapest Convention, States Parties and Reservations, *supra* note 60.

[80] *Id.*

[81] *Id.*

[82] *Id.*

TABLE II: NUMBER OF BUDAPEST CONVENTION RESERVATIONS BY STATE PARTY

Country	Reservations	Declarations	Other[83]	Total
United States of America	6	4		**10**
Japan	4	4		**8**
Switzerland	4	4		**8**
Chile	5	2		**7**
Peru	3	4		**7**
Andorra	5	1		**6**
Israel	6			**6**
Argentina	5			**5**
Azerbaijan	4	1		**5**
Canada	2	3		**5**
Ukraine	2	2	1	**5**
Belgium	2	2		**4**
Denmark	3	1		**4**
France	2	2		**4**
Lithuania	2	2		**4**
Nigeria	4			**4**
Slovak Republic	2	2		**4**
United Kingdom	4			**4**
Australia	3			**3**
Czech Republic	1	2		**3**
Finland	2	1		**3**
Liechtenstein	2	1		**3**
Montenegro	3			**3**
Norway	3			**3**
Sweden	3			**3**

[83] Includes unilateral statements such as understandings and communications.

Country	Reservations	Declarations	Other[83]	Total
Costa Rica		2		2
Georgia		2		2
Germany	1	1		2
Greece	2			2
Hungary	1	1		2
Latvia	2			2
Netherlands		2		2
Sri Lanka	2			2
Austria	1			1
Brazil		1		1
Bulgaria	1			1
Colombia	1			1
Iceland		1		1
Poland	1			1
Portugal		1		1
Republic of Moldova		1		1
Spain		1		1
Total	**94**	**51**	**1**	**146**

2) Which Reservations to Employ?

Unlike the Budapest Convention, the UN Cybercrime Convention does not have a dedicated reservations provision.[84] The VCLT establishes that a State may lodge a reservation unless it is prohibited by a treaty[85]—with no provision expressly prohibiting reservations, one may assume they are permitted. In fact, Argentina and Russia have already indicated their intention to lodge reservations.[86] The VCLT also establishes that a reservation to a provision may not be made if "only specified reservations" are allowed.[87] There is a question of whether the Convention allows

[84] *Compare* UN Cybercrime Convention, *supra* note 2, *with* Budapest Convention, *supra* note 6, art. 42.

[85] VCLT, *supra* note 73, art. 19(a).

[86] *General Assembly: 55th Plenary Meeting (Resumed), 79th Session*, UN WEB TV (Dec. 24, 2024), https://webtv.un.org/en/asset/k1f/k1fyzisiei?kalturaStartTime=3585&kalturaStartTime=3891; Litvishko, *supra* note 12.

[87] VCLT, *supra* note 73, art. 19(b).

reservations only to specific provisions, given that some provisions do explicitly contemplate reservations[88] while others do not. For simplicity, I will assume that reservations are not permitted to provisions that use obligatory language unless so specified. Further research on this question is welcomed.

Adding even more chaos, other articles are not drafted in the language of reservations at all, instead using optional language. For example, on illegal access, "[a] State Party may require … the intent of obtaining electronic data or other dishonest or criminal intent."[89] Nearly 12% (17 of 146) of the reservations lodged under the Budapest Convention were related to intent.[90] But this only affects how a country defines a criminal offense domestically, while the Convention also requires States Parties to "afford one another the widest measure of mutual legal assistance" in both "investigations, prosecutions and judicial proceedings" and for the "collection of evidence in electronic form."[91] A dual criminality reservation, like the 23 (16% of the total) lodged on this topic in the Budapest Convention,[92] is thus also necessary.

Dual criminality, or double criminality, requires that the act that is the object of a request be criminalized both by the requesting State and the requested State.[93] This prevents assisting an investigation into conduct that would not be illegal under a State's own law. Under the UN Cybercrime Convention, dual criminality is merely an optional ground for declining an assistance request.[94] Further, it cannot "be required as a condition for providing" expedited preservation of stored data for the substantive offenses established by the Convention, which may only be declined on the basis of the grounds in Article 40.[95]

Dual criminality may, however, be used as the basis to decline preservation requests with respect to other offenses.[96] This is essential for addressing TNR. Canada, where there is a large Iranian diaspora, has identified the TNR threat the IRI poses as "detrimental to Canada's interests."[97] Yet, when it comes to mutual assistance on criminal matters, Canada does not generally require dual criminality except for seizure and forfeiture orders.[98] Under the Convention, it should reconsider this practice.

[88] *See, e.g.*, UN Cybercrime Convention, *supra* note 2, art. 11(3).

[89] *Id*. art. 7(1–2). See discussion in Section 2.B on intent and the protection of security researchers.

[90] *See* Budapest Convention, States Parties and Reservations, *supra* note 60.

[91] UN Cybercrime Convention, *supra* note 2, art. 40(1).

[92] *See* Budapest Convention, States Parties and Reservations, *supra* note 60.

[93] *See, e.g.*, Peter Rackow & Cornelius Birr, *Recent Developments in Legal Assistance in Criminal Matters*, 2(3) GOETTINGEN J. INT'L L. 1087, 1090–91 (2010).

[94] UN Cybercrime Convention, *supra* note 2, art. 40(8).

[95] *Id*. art. 42(6).

[96] *Id*. art. 42(4–5).

[97] Hogue, *supra* note 44, 94.

[98] *Requesting Mutual Legal Assistance from Canada*, GOV'T CANADA, https://www.justice.gc.ca/eng/cj-jp/emla-eej/mlaguide-guideej.html#:~:text=As%20a%20general%20rule%2C%20dual,always%20required%20under%20Canadian%20law. (last updated July 7, 2021).

While dual criminality is already an obligation for extradition,[99] States can bolster protection through a reservation that modifies their obligations to the same effect as the rejected proposal on an exception for political offenses. They can also add limitations such as a refusal when a person would face trial by an exceptional court or with no legal guarantees, would be likely to serve a sentence under "inhuman conditions," or be subject to the death penalty (for States which have abolished it).[100]

This paper cannot cover all appropriate reservations. States should carefully review the text of the UN Cybercrime Convention to identify the most pressing concerns for them. Additionally, reservations should be consistent across the Budapest and UN frameworks, including for States that may lodge reservations to either instrument for the first time. Finally, States should review the reservations lodged by other States to see if any are appropriate for their domestic systems.[101]

3) Domestic Frameworks

Reservations to the Convention are not the end. Many States will need to enact domestic implementing legislation, and all must abide by the Convention's requirements to criminalize the covered offenses and provide mandatory safeguards.[102] The deference to domestic law, duly criticized by experts,[103] is precisely the basis on which to bolster domestic frameworks.

Even Russia is already assessing its domestic system and considering, for example, a law "aimed at regulating the procedure for ensuring the preservation of electronic data at the request of both foreign and Russian authorities" while simultaneously "preventing Russian providers from fulfilling foreign requests for the preservation or provision of data received directly from abroad."[104] The US has also urged similar action, noting that "implementation of Convention provisions … must be paired with robust domestic safeguards."[105] Large, powerful States should not be the only ones to benefit from these protections.

4. CONCLUSION

I have tried to show some of the risks the UN Cybercrime Convention poses to national security, cybersecurity, and human rights as discussed by industry and

[99] UN Cybercrime Convention, *supra* note 2, art. 37(1).

[100] *See, e.g.*, Treaty Office, *Reservations and Declarations for Treaty No.185 - Convention on Cybercrime (ETS No. 185)*, COUNCIL EUR., https://www.coe.int/en/web/conventions/full-list?module=declarations-by-treaty&numSte=185&codeNature=10&codePays=POR (last visited Jan. 7, 2025) (declaration of Portugal contained in the instrument of ratification deposited on Mar. 24, 2010).

[101] *See* Budapest Convention, States Parties and Reservations, *supr*a note 60.

[102] UN Cybercrime Convention, *supra* note 2, arts. 7–21, 24, 36(2).

[103] Rodriguez, *supra* note 25 (referring to UN Cybercrime Convention, *supra* note 2, art. 36(1)(a)).

[104] Litvishko, *supra* note 12.

[105] *Explanation of Position of the United States on the Adoption of the Resolution on the UN Convention Against Cybercrime, supra* note 32.

rights experts alike. The origins of the Convention, as well as its relationship to the Budapest Convention, are essential for States to assess whether to ratify either or both frameworks. There are undoubtedly competing visions of cyber governance at play. Whether States choose to opt in or out, they should adopt the strategies laid out herein, such as making use of reservations and bolstering domestic frameworks to best protect security interests and human rights—two categories that are deeply intertwined in this context.

Every country must thoroughly assess its goals, risk positioning, and capacity regarding the UN Cybercrime Convention and whether it is appropriate to opt in or out. Many non-Western countries have joined the Budapest Convention framework without employing the tools their European and high-income counterparts have used. This pattern should not be replicated under the UN Cybercrime Convention. It is not too late for any country considering either instrument to protect themselves and their populations while effectively countering cybercrime.

ACKNOWLEDGMENTS

I would like to thank the reviewers and my Atlantic Council colleagues, especially Emma Schroeder, for their thoughtful comments and feedback. All errors are my own.

CyCon 2025: The Next Step
17th International Conference on Cyber Conflict
C. Kwan, N. Gratzer, K. Podiņš, M. Tolppa (Eds.)
2025 © NATO CCDCOE Publications, Tallinn

Permission to make digital or hard copies of this publication for internal use within NATO and for personal or educational use when for non-profit or non-commercial purposes is granted provided that copies bear this notice and a full citation on the first page. Any other reproduction or transmission requires prior written permission by NATO CCDCOE.

CyCon 2025: Responding to Ransomware Within the Boundaries of International Law

Tsvetelina J. van Benthem
Research Fellow
Oxford Institute for Ethics, Law and Armed Conflict
Blavatnik School of Government
University of Oxford
Oxford, United Kingdom
tsvetelina.vanbenthem@bsg.ox.ac.uk

Roxana Radu
Associate Professor of Digital Technologies and Public Policy and Hugh Price Fellow
Blavatnik School of Government and Jesus College
University of Oxford
Oxford, United Kingdom
roxana.radu@bsg.ox.ac.uk

Abstract: This paper explores the interaction between state obligations under international law and the domestic measures taken by states to counter the ransomware threat. On the one hand, states are required to take a proactive approach by taking steps to protect against ransomware operations. On the other hand, their freedom to take action against ransomware actors and related harms is not unlimited – obligations under international law, such as those emanating from state sovereignty, constrain state action in important ways. Understanding the boundaries of action compliant with international law is essential: a balance must be struck between pursuing effective ransomware responses and ensuring compliance with international law. Maintaining trust in the international legal system is contingent upon clearly signalled and honoured state commitments to international law – its substance, institutions and processes.*

Keywords: *international law, offensive cyber operations, positive obligations, ransomware, resilience-building*

* This publication arises from research funded by the John Fell Oxford University Press Research Fund (grant no 14333).

1. INTRODUCTION

Ransomware has established itself as one of the most pervasive and disruptive contemporary threats.[1] According to a 2024 report from the European Union Agency for Cybersecurity, the ransomware threat is characterized by 'ongoing growth' and a changing toolbox of extortion tactics.[2] Experience from past years has shown that effective protection can only be ensured by continuously evolving counter-ransomware strategies, establishing diverse and comprehensive resilience and disruptive measures, and streamlining collective responses.[3]

As the ransomware threat landscape continues to evolve, so do the domestic policies of states to protect themselves and those under their jurisdictions from its criminal ecosystem. States have adopted a range of measures, individually and collectively, to build domestic resilience, criminalize the deployment of ransomware, bolster law enforcement capabilities and, in some cases, offensively disrupt ransomware networks. At the same time, states continue to signal their commitment to the rules-based international order and the important role that international law plays in countering harms produced via information and communications technologies (ICTs).[4] This paper explores the interaction between state obligations under international law and the domestic measures taken by states to counter the ransomware threat. On the one hand, states are required to take a proactive approach to protect against ransomware operations. On the other hand, their freedom to take action against ransomware actors and related harms is not unlimited – obligations under international law constrain state action in important ways. Understanding the boundaries of action compliant with international law is essential: a balance must be struck between pursuing effective ransomware responses and ensuring compliance with international law. Maintaining trust in the international legal system is contingent upon clearly signalled and honoured state commitments to international law – its substance, institutions and processes.

The paper proceeds in four sections. Section 2 lays the foundation for the analysis by defining ransomware and exploring key trends in the evolution of the ransomware ecosystem. Section 3 addresses the interaction between ransomware, domestic responses and international law-making. Section 4 turns to the positive and

[1] Microsoft, 'Digital Defense Report 2024: The Foundations and New Frontiers of Cybersecurity' 27 <https://www.microsoft.com/en-us/security/security-insider/intelligence-reports/microsoft-digital-defense-report-2024> accessed 12 April 2025.

[2] European Union Agency for Cybersecurity, 'ENISA Threat Landscape 2024' (September 2024) 45 <https://www.enisa.europa.eu/sites/default/files/2024-11/ENISA%20Threat%20Landscape%202024_0.pdf> accessed 12 April 2025).

[3] Interpol, Australian Government and International Counter-Ransomware Task Force, 'A Comparative Threat Assessment on Counter-Ransomware Interventions' (September 2024).

[4] UNGA, 'Report of the Open-Ended Working Group on Security of and in the Use of Information and Communications Technologies 2021–2025' (22 July 2024) UN Doc A/79/214, para 35.

negative obligations binding states under international law in their development and implementation of counter-ransomware measures and applies these obligations to concrete examples of measures taken or signalled by states. Section 5 concludes.

2. RANSOMWARE: DEFINITION AND TRENDS

Ransomware is, at base, a form of malware designed to take control of a target's assets, with assets rendered unavailable until a demand is met.[5] While the most prevalent earlier form of ransomware involved the encryption of data on the target's device, with decryption being contingent on the payment of a ransom, current trends suggest the increasing use of exfiltration of data from the victim's device coupled with threats to publish or sell sensitive data in case of non-payment of the ransom.[6]

In recent years, the ransomware model has become steadily more sophisticated and professionalized. A ransomware-as-a-service ecosystem continues to proliferate, with emerging platforms such as RansomHub and Farnetwork.[7] As Bátrla and Harašta explain, this ecosystem contributes

> to the development of [a] complex, diverse, and market-forces driven system comprising interactions between specialized actors, such as malware developers and operators, affiliates, analysts, botmasters, initial access merchants, money processing and laundering specialists, escrow services, forum and illicit marketplace administrators, infrastructure administrators, [and] even negotiation and customer support personnel.[8]

What this means is that any action to counter ransomware must face an entire criminal system. This criminal system is predominantly composed of private actors. That being said, private actors exhibit varying proximity and coordination with state actors. Most often, ransomware groups operate from safe-haven jurisdictions that are unwilling to take decisive steps to dismantle them. This, in turn, impedes traditional enforcement action.[9] In some cases, these groups are themselves sponsored by a

[5] ENISA (n 2) 45; 'Oxford Statement on International Law Protections in Cyberspace: The Regulation of Ransomware Operations' (*The Oxford Process*) <https://www.elac.ox.ac.uk/the-oxford-process/the-statements-overview/the-oxford-statement-on-ransomware-operations/#:~:text=3.-,States%20must%20refrain%20from%20conducting%2C%20directing%2C%20authorising%20or%20aiding%20and,and%20opinion%2C%20freedom%20of%20expression%2C> accessed 12 April 2025.

[6] UK House of Commons and House of Lords, Joint Committee on the National Security Strategy, 'A Hostage to Fortune: Ransomware and UK National Security', First Report of Session 2023–2024 (13 December 2023) 5.

[7] ENISA (n 2) 30.

[8] M Bátrla and J Harašta, '"Releasing the Hounds?" Disruption of the Ransomware Ecosystem Through Offensive Cyber Operations' (2022) *Proceedings of the 14th International Conference on Cyber Conflict* 96.

[9] Ransomware Task Force, 'Combating Ransomware: A Comprehensive Framework for Action: Key Recommendations from the Ransomware Task Force' (2021) 27.

state.[10] In others, the activities of the criminal groups – disrupting and destroying foreign interests and assets – align with the goals of the territorial jurisdiction and are therefore tolerated.[11] Geopolitical events expose divergent alignments between states and ransomware groups – for instance, the cybercriminal group Conti pledged support for Russia in its war against Ukraine.[12] Depending on the modality of the private actor–state relationship, the actions of the former may be attributable to the latter.[13] Exploring the content of the customary rules of attribution and their application to various ransomware groups is beyond the scope of this paper.[14] For present purposes, it is important that ransomware operations from safe-haven jurisdictions significantly hamper efforts to bring the law to bear on perpetrators. This, in turn, sheds light on the importance of international cooperation in criminal matters, including in law enforcement.

Finally, while artificial intelligence (AI) can be used in the fight against ransomware groups – in tracking threat actors, scenario planning and resilience-building[15] – it is equally employed by ransomware actors to facilitate their activities. It is projected that AI will increase the scale and impact of ransomware operations, especially in relation to reconnaissance and social engineering.[16]

Against this background, states continue to tailor their approaches to the evolving ransomware ecosystem, increasingly focusing on domestic resilience-building and cooperation in the dismantling of ransomware criminality.

[10] For evidence that North Korea sponsored The Lazarus Group, a hacking team behind the WannaCry 2.0 global ransomware attack, see eg US Department of Justice, 'North Korean Regime-Backed Programmer Charged With Conspiracy to Conduct Multiple Cyber Attacks and Intrusions' (Press Release, 6 September 2018) <https://www.justice.gov/opa/pr/north-korean-regime-backed-programmer-charged-conspiracy-conduct-multiple-cyber-attacks-and> accessed 22 April 2025.

[11] For a discussion of the Kremlin's approach to certain ransomware actors in the United Kingdom, see House of Commons and House of Lords, Joint Committee on the National Security Strategy (n 6) 17–18.

[12] Global Initiative against Transnational Organized Crime, 'The Rise and Fall of the Conti Ransomware Group' (27 June 2023) <https://globalinitiative.net/analysis/conti-ransomware-group-cybercrime/> accessed 12 April 2025.

[13] The recognized customary grounds of attribution under international law are codified in International Law Commission, Articles on the Responsibility of States for Internationally Wrongful Acts (2001) arts 4–11. Of most relevance are the grounds under arts 4 (most relevantly on *de facto* organs), 8 (instructions, direction or control) and 11 (acknowledgement and adoption).

[14] The following sections will focus on the scenario of criminal groups whose conduct is not legally attributable to a particular state or states.

[15] International Counter Ransomware Initiative, '2024 Joint Statement' (2 October 2024) <https://www.whitehouse.gov/briefing-room/statements-releases/2024/10/02/international-counter-ransomware-initiative-2024-joint-statement/> accessed 10 April 2025; S Poudyal and D Dasgupta, 'AI-Powered Ransomware Detection Framework' (2020) *IEEE Symposium Series on Computational Intelligence*.

[16] UK National Cyber Security Centre, 'The Near-Term Impact of AI on the Cyber Threat' (2024) <https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat#section_3> accessed 12 April 2025.

3.THE INTERACTION BETWEEN RANSOMWARE, DOMESTIC MEASURES AND INTERNATIONAL LAW-MAKING

That ransomware poses a transnational threat to lives and livelihoods, the global economy, and the normal functioning of governments and the private sector is, by now, both undeniable and universally understood. International efforts, such as the International Counter-Ransomware Initiative,[17] seek to pool knowledge, build collective resilience and develop common policies to counter the threat. In November 2024, the United Nations Security Council heard briefings on the challenges posed by ransomware attacks against hospitals and other healthcare facilities and services,[18] with the United States representative stating that 'none of us is doing enough'.[19]

How states respond to ransomware individually and collectively inevitably involves international law. On the one hand, their responses demonstrate the connection between ransomware harms and human rights, including the rights to life and health. Ransomware operations pose foreseeable risks to the enjoyment of these rights: 'Health experts have estimated that ransomware attacks were responsible for the deaths of dozens of patients in the United States.'[20] On the other hand, international law constrains state responses both within their territory (for instance, in taking measures that do not violate the human rights of those under their jurisdiction) and extraterritorially (for instance, through obligations that protect the interests of other states, such as sovereignty, non-intervention and the prohibition of the use of force, and the rights of individuals). International law is both an important item in the state toolkit for combating ransomware[21] and an important reminder that their freedom to take action against ransomware actors is not unlimited.

There are three notable dynamics in the interaction between ransomware, domestic measures and international law-making.

First, while ransomware is clearly a threat of utmost concern to states, their statements on the application of international law to cyberspace do not suggest any difference in

[17] The International Counter-Ransomware Initiative is an initiative uniting more than 70 member states and organizations to build cross-border resilience and collectively disrupt and defend against cyber actors; see 'About' (*International Counter-Ransomware*) <https://counter-ransomware.org/> accessed 13 April 2025.

[18] WHO, 'Director-General's Remarks at Meeting of the UN Security Council on Threats Posed by Ransomware Attacks Against Hospitals and Other Health-Care Facilities and Services' <https://www.who.int/director-general/speeches/detail/who-director-general-s-remarks-at-meeting-of-the-un-security-council-on-threats-posed-by-ransomware-attacks> accessed 8 April 2025.

[19] US Mission to the UN, 'Remarks at a UN Security Council Briefing on Ransomware Attacks Against Hospitals and Other Healthcare Facilities and Services' <https://usun.usmission.gov/remarks-at-a-un-security-council-briefing-on-ransomware-attacks-against-hospitals-and-other-healthcare-facilities-and-services/> accessed 10 April 2025.

[20] ibid.

[21] T van Benthem and C Tams, 'Regulating Ransomware Through International Law' (2024) Report of the Scottish Council on Global Affairs <https://scga.scot/wp-content/uploads/2024/02/Ransomware-Report-Final-January-2024.pdf> accessed 10 April 2025.

the specification of international law obligations in their application to ransomware. The approach taken is generic, with specific types of cyber operations often given merely as illustrations. Ransomware operations are used as illustrations in the 2024 position of Austria (exemplifying the trigger for a positive due diligence obligation),[22] the 2024 position of the Czech Republic (exemplifying breaches of sovereignty)[23] and the 2023 position of Costa Rica (exemplifying breaches of sovereignty through loss of functionality in operating systems, intervention and due diligence, and the meaning of attack under international humanitarian law).[24] While certain legal interpretations may be seen as implicitly tied to the ransomware threat,[25] national positions do not explicitly suggest that ransomware operations are shifting or specifying their interpretations of the law.

Second, the development of state positions and their content is bound to states' perceptions of their own vulnerability and their technical capacities to counter the threat. For instance, Costa Rica's national position was adopted in the aftermath of a large-scale and disruptive ransomware attack against the state, and its position paper highlights that ransomware 'may have significant economic, political, and human costs, as the ransomware attacks targeting Costa Rica in 2022 illustrates'.[26] And more limited interpretations of sovereignty-related rules of international law may be connected to the ability and willingness of states to counter ransomware groups extraterritorially through disruptive operations. Though not specifically stated in national positions, the need to protect against and respond to ransomware inevitably shapes the legal positions that states advance internationally.

Third, the changing ransomware landscape, perceived necessity of responses and domestic action may lead to an evolution in the international legal rules, both through development in customary law and evolving interpretations of relevant treaties. For instance, perceived needs to act extraterritorially by accessing criminal infrastructure without the consent of the territorial state may invite a rethinking of the content of sovereignty, non-intervention and the use of force, together with related justifications in primary rules and circumstances precluding wrongfulness under the law of state responsibility. The very methodology of customary international law formation[27] and

[22] Republic of Austria, 'Cyber Activities and International Law' Position Paper (April 2024) 11.

[23] Czech Republic, 'Position Paper on the Application of International Law in Cyberspace' (2024) para 6(c).

[24] Costa Rica, 'Position on the Application of International Law in Cyberspace' (2023) paras 20, 25, 28, 49.

[25] For instance, Denmark, 'Denmark's Position Paper on the Application of International Law in Cyberspace' (2023) suggests that 'a cyber operation resulting in the malfunctioning of a State's financial system leads to significant economic damage' may fall within the purview of the prohibition of the use of force under the Charter of the United Nations.

[26] Costa Rica, 'Position on the Application of International Law in Cyberspace' (2023) para 3.

[27] International Law Commission, 'Draft Conclusions on the Identification of Customary International Law' (2018), with the requirements of (1) practice that is widespread, representative and consistent and (2) acceptance of such practice as law.

the rules on the interpretation of treaties[28] are important bulwarks against expansive interpretations and developments originating in a minority of states.

The following section examines the interaction between domestic measures taken or signalled by states, and their relation to positive and negative obligations under international law.

4. APPLYING INTERNATIONAL LAW TO STATE MEASURES IN COUNTERING RANSOMWARE

States have consistently affirmed that international law applies to the use of ICTs.[29] And while ransomware is not regulated under international law as such, a range of international obligations of general application are relevant to its regulation. Thus, under international law, states are bound by positive and negative obligations in relation to, among others, individuals and other states, and these general obligations apply when states take measures to counter ransomware activity. The following sections examine, first, positive obligations that require states to take steps to tackle the ransomware threat, and second, a subset of negative obligations that constrain states in the measures they are allowed to take.

A. Obligations to Take Steps: Positive Due Diligence Obligations to Counter Ransomware

Ransomware's impact is felt across society. Its harmful effects on the provision of essential services, the operation of governmental entities and the private sector, and the security and well-being of individuals are well-documented.[30] The threat of ransomware is both real and foreseeable. Unsurprisingly, there is increasing state activity in designing domestic strategies to counter ransomware, in coordinating with local and international partners, and in building or enhancing the existing regulatory framework. While this state activity is in line with states' interests, and a reflection of their respective perceptions of vulnerability, the taking of measures to counter the ransomware threat is also a matter of international obligation.

Under international law, positive obligations compel states into action. Such positive obligations exist under different legal regimes and under different sources of law. For

[28] See also International Law Commission, 'Draft Conclusions on Subsequent Agreements and Subsequent Practice in Relation to the Interpretation of Treaties' (2018).

[29] UNGA, 'Final Substantive Report of the Open-ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security' (10 March 2021) UN Doc A/AC.290/2021/CRP.2 para 34; 'Report of the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security' (14 July 2021) UN Doc A/76/135, paras 69ff.

[30] J MacColl, P Hüsch, G Mott, J Sullivan, JRC Nurse, S Turner and N Pattnaik, 'The Scourge of Ransomware: Victim Insights on Harms to Individuals, Organisations and Society' (2024) RUSI Occasional Paper.

instance, international human rights law treaty instruments[31] and customary human rights contain positive obligations binding states, triggered in cases of foreseeable risks to particular rights of individuals under their jurisdiction. Where such foreseeable risks arise, states must take steps to prevent their materialization or mitigate their effects, including in cases where the risks are created by non-state actors. Elaborating on this obligation in the context of the right to life under the International Covenant on Civil and Political Rights, the United Nations Human Rights Committee explained that

> State parties are thus under a due diligence obligation to take reasonable, positive measures that do not impose disproportionate burdens on them in response to reasonably foreseeable threats to life originating from private persons and entities whose conduct is not attributable to the State.[32]

A similar interpretation of the right to life was given by the African Commission on Human and People's Rights: 'The State has a positive duty to protect individuals and groups from real and immediate risks to their lives caused either by actions or inactions of third parties.'[33]

Positive obligations arise under a wide range of rights, including life, health, privacy, property and education, and there is no prescriptive list of measures that must be implemented to discharge them in each and every case and across contexts. When it comes to countering ransomware, these obligations may be discharged by a number of technical, legal or institutional measures, such as the enactment of domestic legislation to criminalize ransomware and impose cybersecurity requirements on local entities, the adoption of measures and establishment of government structures to prevent ransomware operations or halt ongoing ones, the drafting and publicizing of contingency plans and cyber hygiene, the investigation and prosecution of those responsible, the adoption of guidance on ransomware payments, among many others.[34] The Human Rights Committee has previously stated that states should develop, 'when necessary, contingency plans and disaster management plans designed to increase preparedness' in view of 'massive cyberattacks resulting in disruption of essential services'.[35] It bears mentioning that the United Nations Convention on Cybercrime,

[31] For instance, under the International Covenant on Civil and Political Rights (16 December 1966) 999 UNTS 171; African Charter on Human and Peoples' Rights (21 October 1986) 1520 UNTS 217; American Convention on Human Rights (22 November 1969) 1144 UNTS 123; European Convention on Human Rights (4 November 1950) 213 UNTS 222.

[32] Human Rights Committee, 'General Comment 36 on the Right to Life' (2018) para 21.

[33] African Commission on Human and Peoples' Rights, 'General Comment No 3 on the Right to Life' (2015) para 41.

[34] For examples of such measures, see 'Oxford Statement on International Law Protections in Cyberspace: The Regulation of Ransomware Operations' (*The Oxford Process*) <https://www.elac.ox.ac.uk/the-oxford-process/the-statements-overview/the-oxford-statement-on-ransomware-operations/#:~:text=States%20must%20refrain%20from%20conducting,Charter%20of%20the%20United%20Nations> accessed 12 April 2025.

[35] Human Rights Committee, 'General Comment 36 on the Right to Life' (2018) para 26.

adopted by the United Nations General Assembly in December 2024 and open for signature in 2025, would impose a number of substantive criminalization obligations and jurisdictional, enforcement and institutional ones on its parties. Although the Convention does not specifically criminalize ransomware, the offences of illegal access, illegal interception, interference with electronic data and misuse of devices, among others, would cover such conduct.[36] In this way, compliance with substantive and procedural obligations under the Cybercrime Convention could align with the demands of positive obligations under international human rights law.

Another example of an international law obligation requiring states to take steps is the obligation for states to not knowingly allow their territory to be used for acts contrary to the rights of other states.[37] This obligation, expounded on by the International Court of Justice in the *Corfu Channel* case, is also characterized by a due diligence standard, and is aimed at the protection of the interests of states.[38] While there are ongoing controversies over the customary scope of this rule,[39] it undeniably has important implications for ransomware operations. A state's failure to dismantle or otherwise counter ransomware actors under its jurisdiction who are conducting ransomware operations affecting other states, where the state of jurisdiction is or should be aware of their activity, would result in a breach of this rule and thereby entail the responsibility of the state. This rule therefore has the capacity to provide redress against states that have become safe havens for ransomware criminal groups.

While these positive obligations are flexible and subject to state capacity, they clearly demonstrate that, where their triggering circumstances are met, states are required to be proactive. In the context of ransomware, many states are indeed taking a proactive approach to building resilience against cyber threats, including ransomware. Australia, for instance, has introduced mandatory reporting of cybersecurity incidents for critical infrastructure operators under the Security of Critical Infrastructure Act 2018.[40] The Cyber Security Strategy Action Plan 2023–2030 stresses the importance of building resilience within society, providing clear guidelines for businesses, and creating a comprehensive threat intelligence framework.[41] The United Kingdom, in its '2023 Ransomware White Paper', adopted a holistic and systemic approach to ransomware,

[36] UNGA, 'United Nations Convention Against Cybercrime; Strengthening International Cooperation for Combating Certain Crimes Committed by Means of Information and Communications Technology Systems and for the Sharing of Evidence in Electronic Form of Serious Crimes' (31 December 2024) UN Doc A/RES/79/243 arts 7, 8, 9, 11.

[37] *Corfu Channel (UK v Albania)* [1949] ICJ Rep 4 [22].

[38] A Coco and T de Souza Dias, 'Cyber Due Diligence: A Patchwork of Protective Obligations in International Law' (2021) 32(3) European Journal of International Law 771.

[39] 'Due Diligence' (*CyberLaw Toolkit, CCDCOE*) <https://cyberlaw.ccdcoe.org/wiki/Due_diligence> accessed 8 April 2025.

[40] Australian Government, 'Security of Critical Infrastructure Act 2018 (SOCI)' (*Federal Register of Legislation*) <https://www.legislation.gov.au/C2018A00029/latest/versions> accessed 8 April 2025.

[41] Australian Government, 'Cyber Security Strategy Action Plan 2023–2030' 6, 8, 14 <https://www.homeaffairs.gov.au/cyber-security-subsite/files/2023-cyber-security-strategy-action-plan.pdf> accessed 8 April 2025.

with a particular focus on cyber hygiene.[42] And Costa Rica, following the highly disruptive ransomware operation affecting the country in 2022,[43] has focused on the protection of its critical infrastructure with periodic analyses of vulnerability and risk,[44] and the bolstering of entities tasked with cybersecurity coordination, such as the Centro de Respuesta de Incidentes de Seguridad Informatica. Importantly, Costa Rica has emphasized that measures meant to tackle cybersecurity threats must be undertaken in compliance with human rights, especially freedom of expression and privacy.[45] Specific incidents, notably the ransomware operation against Colonial Pipeline in the United States, have also prompted tailored responses, such as measures to protect the security of supply chains, the development of playbooks for responding to cybersecurity incidents and the establishment of better evidence-sharing arrangements between the government and the private sector.[46]

Effectively discharging positive obligations under international law could also require transnational coordination and cooperation, as individual states may be unable to protect against the ransomware threat on their own. Participation in transnational collaborative initiatives, such as CyberSouth+, jointly launched by the European Union and Council of Europe,[47] can enhance collaborative processes, including by strengthening the tools of criminal justice on the disclosure of electronic evidence.

It can be concluded that states are obliged under international law to take measures to protect individuals and other states from the harmful effects of ransomware. At the same time, the freedom of states to take such measures is not unlimited. The boundaries of their freedom are determined by a number of negative obligations under international law.

B. Limited Freedom: Obligations to Abstain from Particular Types of Measures While Countering Ransomware

While positive obligations require states to act, a range of negative obligations under international law constrain the freedom of states to take these measures. As discussed in Section 2, most ransomware actors operate from safe-haven jurisdictions

[42] UK National Cyber Security Centre and National Crime Agency, 'Ransomware, Extortion and the Cyber Crime Ecosystem' (2023) White Paper <https://www.ncsc.gov.uk/files/White-paper-Ransomware-extortion-and-the-cyber-crime-ecosystem.pdf> accessed 8 April 2025.

[43] 'Costa Rica Ransomware Attack' (*CyberLaw Toolkit, CCDCOE*, 2022) <https://cyberlaw.ccdcoe.org/wiki/Costa_Rica_ransomware_attack_(2022)> accessed 8 April 2025.

[44] R García Villalobos and others, 'Protocolo para el desarrollo de acciones que se deben implementar ante una amenaza de un ataque a la ciberseguridad nacional' (2022).

[45] Costa Rica, 'Estrategia Nacional de Ciberseguridad' (2017) 8 <https://www.clubdeinvestigacion.com/wp-content/uploads/2022/11/Estrategia-Nacional-de-Ciberseguridad-Costa-Rica-2022.pdf> accessed 8 April 2025.

[46] Kimberly Wood, 'Cybersecurity Policy Responses to the Colonial Pipeline Ransomware Attack' (2023) (*Georgetown Environmental Law Review*, 7 March 2023) <https://www.law.georgetown.edu/environmental-law-review/blog/cybersecurity-policy-responses-to-the-colonial-pipeline-ransomware-attack> accessed 8 April 2025.

[47] This initiative seeks to entrench collaboration within the framework of the Budapest Convention on Cybercrime with Algeria, Egypt, Jordan, Lebanon, Libya, Morocco, Palestine and Tunisia, available at https://www.coe.int/en/web/cybercrime/cybersouthplus.

beyond the reach of law enforcement authorities of target states. Rules based on the principle of state sovereignty, such as the prohibition on the use of force, the principle of non-intervention and the primary rule of sovereignty, all constrain extraterritorial enforcement activities without the consent of the state in whose territory the enforcement operation is to take place.[48] Since in most cases such consent will not be forthcoming, the capacity of states to enforce their laws against ransomware actors will be limited by international law.

At the same time, states signal an interest in a proactive 'offensive' approach to the dismantling of ransomware groups and cyber threats more generally. Australia, for instance, has invested heavily in expanding the range and sophistication of its offensive and defensive cyber capabilities. The Australian Signals Directorate undertakes offensive cyber operations to support national security, with one of its functions being to prevent and disrupt offshore cyber-enabled crime.[49] In a 2018 speech, the Director-General of the Australian Signals Directorate explained that its activities focused on offshore use 'specialized tools and techniques to disrupt their [their adversaries'] communications or interfere with the way they operate online'.[50] And in the United Kingdom, the report of the House of Commons and House of Lords Joint Committee on the National Security Strategy recommended that the Government 'invest significantly more resources in the National Crime Agency's response to ransomware, enabling it to pursue a more aggressive approach to infiltrating and disrupting ransomware operators'.[51]

While a more aggressive extraterritorial approach may *prima facie* seem an effective way of tackling the ransomware threat, it would come into tension with a range of negative obligations under international law that constrain extraterritorial enforcement activities.[52] What complicates the analysis under these negative obligations are the ongoing controversies over their elements and, for some, their very existence as primary rules of international law. Even if a particular extraterritorial activity would constitute a violation of a particular negative obligation, it may still – depending on the obligation breached – be possible to resort to justifications, either in the primary rules themselves or under the customary law of state responsibility. The analysis first

[48] International human rights law also imposes constraints on extraterritorial state conduct. For reasons of scope, this strand of analysis is not addressed in this paper.

[49] Australian Government, Australian Signals Directorate, 'ASD Corporate Plan 2023–24' <https://www.asd.gov.au/about/accountability-governance/publications/asd-corporate-plan-2023-24> accessed 9 April 2025.

[50] Australian Government, Australian Signals Directorate, 'Director-General ASD Speech to the Lowy Institute' <https://www.asd.gov.au/news-events-speeches/speeches/director-general-asd-speech-lowy-institute> accessed 9 April 2025.

[51] UK House of Commons and House of Lords, Joint Committee on the National Security Strategy, 'A Hostage to Fortune: Ransomware and UK National Security, First Report of Session 2023–2024' (13 December 2023) 66.

[52] T van Benthem, J Kulesza, Y Liu and N Sun, 'Jurisdiction in Cyberspace' (2024) Sino-European Expert Working Group on the Application of International Law in Cyberspace (EWG-IL), Research Group Report 2024 <https://www.gcsp.ch/sites/default/files/2024-12/EWG-IL_Partnered_Jurisdiction_2024-11%3Bdigital.pdf> accessed 9 April 2025.

turns to the relevant primary obligations before reviewing the possibility of resorting to justifications.

To begin with, as explained by the Permanent Court of International Justice in the Lotus case, under customary law, a state 'may not exercise its power in any form in the territory of another State' without a permissive rule to the contrary.[53] A lawful exercise of extraterritorial enforcement jurisdiction would depend on 'valid consent by a foreign government to exercise jurisdiction on its territory' or 'a specific allocation of authority under international law'.[54] If extraterritorial cyber operations to disrupt ransomware groups qualify as enforcement actions, failing the existence of a permissive ground, they would come into tension with this customary rule.

Further, under the Charter of the United Nations, 'all Members shall refrain in their international relations from the threat or use of force against the territorial integrity or political independence of any state, or in any other manner inconsistent with the Purposes of the United Nations'.[55] The key interpretative question here is over the meaning of 'force', in particular the types of effects and gravity sufficient to qualify as 'force'. Australia's position suggests that '[i]n determining whether a cyber activity constitutes a use of force, States should consider whether the activity's scale and effects are comparable to traditional kinetic operations that rise to the level of use of force under international law', and this entails an analysis of the 'intended or reasonably expected direct and indirect consequences of the cyber activity, including for example whether the activity could reasonably be expected to cause serious or extensive ("scale") damage or destruction ("effects") to life, or injury or death to persons, or result in damage to the victim State's objects, critical infrastructure and/or functioning'.[56] According to some states, including France, cyber operations without *physical* effects may also, depending on the circumstances, be characterized as a use of force.[57] An extraterritorial operation against a ransomware group causing effects in the territory of another state may therefore amount to a use of force under this prohibition.

Beyond the use of force, the principle of non-intervention prohibits states from coercive interferences in the *domaine réservé* of other states.[58] As for the prohibition of the use of force, the contours of this obligation remain contested, in particular regarding the element of coercion. Does coercion imply effects on the 'will' of the other state, or on its 'ability to control its sovereign choices'?[59] The United Kingdom

[53] *SS Lotus (France v Turkey)* [1927] PCIJ Series A No 10, [45].

[54] Michael N Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (CUP 2017) rule 11.

[55] Charter of the United Nations art 2(4).

[56] Australian Government, 'Australia's Submission on International Law to Be Annexed to the Report of the 2021 Group of Governmental Experts on Cyber' 2.

[57] French Ministry of the Armies, 'International Law Applied to Operations in Cyberspace' (2019) 7.

[58] *Military and Paramilitary Activities in and Against Nicaragua (Nicaragua v US)* [1986] ICJ Rep [202].

[59] Marko Milanovic, 'Revisiting Coercion as an Element of Prohibited Intervention in International Law' (2023) 117(4) *American Journal of International Law* 601, 626–48.

seems to adopt a wider understanding of ‘coercion’, explaining that ‘an intervention in the affairs of another State will be unlawful if it is forcible, dictatorial, or otherwise coercive, depriving a State of its freedom of control over matters which it is permitted to decide freely by the principle of State sovereignty’.[60] Under a wider interpretation of the element of coercion, an extraterritorial operation to dismantle a non-state ransomware criminal group may indeed be seen as depriving the territorial state of control over enforcement activities in its jurisdiction.

And finally, while it is uncontested that sovereignty is a principle of international law animating a number of primary rules, there are ongoing debates over its existence and content as a self-standing rule. In their national positions on the application of international law to cyberspace, states increasingly adopt the sovereignty-as-a-rule approach.[61] The United Kingdom, however, has consistently rejected this view.[62] Depending on how a primary rule of sovereignty is framed, it can be more or less constraining on states that seek to counter ransomware actors extraterritorially. The African Union, for example, adopts a wide approach to sovereignty, which would capture any unauthorized access:

> By virtue of territorial sovereignty, any unauthorized access by a State into the ICT infrastructure located on the territory of a foreign State is unlawful. Therefore, the African Union emphasizes that the obligation to respect the

[60] Attorney General, the Rt Hon Suella Braverman QC MP, ‘International Law in Future Frontiers’ (*GOV.UK*, 2022) <https://www.gov.uk/government/speeches/international-law-in-future-frontiers> accessed 9 April 2025.

[61] The following positions accept that sovereignty is a standalone rule of international law: African Union, ‘Common African Position on the Application of International Law to the Use of Information and Communication Technologies in Cyberspace’ (February 2024) para 13; Republic of Austria (n 22) 4; Brazil national position in *GGE 2021 Official Compendium* 18 (Brazil argues that any exception to the rule of sovereignty would require broad state practice and sufficient *opinio iuris*); Government of Canada, ‘International Law Applicable in Cyberspace’ (2022) para 13; China, ‘Views on the Application of the Principle of Sovereignty in Cyberspace’ 2–3; Costa Rica, ‘Position on the Application of International Law in Cyberspace’ (2023) para 19; Czech Republic, ‘Position Paper on the Application of International Law in Cyberspace’ (2024) para 3; Denmark (n 25) 448; Estonia, Contribution to *GGE 2021 Official Compendium* 25; Finland, ‘International Law and Cyberspace’, National Position, (2020) 2–3; France, ‘International Law Applied to Operations in Cyberspace’, paper shared by France with the open-ended working group established by Resolution 75/240, 3; German Federal Government, ‘On the Application of International Law in Cyberspace’ Position Paper (2021) 3; Ireland, ‘Position Paper on the Application of International Law in Cyberspace’ (2023) para 5; Italy, ‘International Law and Cyberspace’, Position Paper, 4; Ministry of Foreign Affairs of Japan, ‘Basic Position of the Government of Japan on International Law Applicable to Cyber Operations’ (28 May 2021) 2; Government of the Kingdom of the Netherlands, ‘Appendix: International Law in Cyberspace’ (2019) 2; New Zealand, ‘The Application of International Law to State Activity in Cyberspace’ (2020) para 12; Norway, ‘Norwegian Positions on Selected Questions of International Law Relating to Cyberspace’ (2021); Poland, ‘The Republic of Poland’s Position on the Application of International Law in Cyberspace’ (2022) 3; Romania, Contribution to *GGE 2021 Official Compendium* 76; Singapore, Contribution to *GGE 2021 Official Compendium* 83; Switzerland, Contribution to *GGE 2021 Official Compendium* 87.

[62] Attorney General’s Office and The Rt Hon Suella Braverman KC MP, ‘International Law in Future Frontiers’ (GOV.UK, 2022) https://www.gov.uk/government/speeches/international-law-in-future-frontiers (accessed 9 April 2025). ‘The general concept of sovereignty by itself does not provide a sufficient or clear basis for extrapolating a specific rule of sovereignty or additional prohibition for cyber conduct going beyond that of non-intervention.’

> territorial sovereignty of States, as it applies in cyberspace, does not include a de minimis threshold of harmful effects below which an unauthorized access by a State into the ICT infrastructure located on the territory of a foreign State would not be unlawful.[63]

While most states condition this rule through a *de minimis* approach regarding effects, the ongoing uncertainty over the content of this rule creates a significant grey area regarding the legality of extraterritorial measures to tackle ransomware groups where the consent of the territorial state has not been obtained.

Importantly, even if a state is in breach of its international obligations when conducting extraterritorial counter-ransomware activities, this is not the end of the analysis. The state may be able to rely on justifications. For instance, states can use force in self-defence if they become the victim of an armed attack. Under the traditional restrictive view of the content of self-defence, it must be determined whether a ransomware operation that amounts to an armed attack actually occurred and whether the actor initiating that attack was a state.[64]

States may be able to rely on justifications under the law of state responsibility, such as countermeasures and necessity. Countermeasures are measures that, but for the internationally wrongful act of the responsible state, would be contrary to the international obligations of the state undertaking the measure. It is the fact that they respond to a prior illegality that provides the basis for their justifiability. The measures must comply with a number of stringent requirements related to their purpose and proportionality, among others, and must not affect a number of foundational obligations of international law, including the obligation to refrain from the threat or use of force as embodied in the Charter of the United Nations and obligations for the protection of fundamental human rights.[65] Importantly for counter-ransomware operations, states have a basis to resort to countermeasures not only against states that themselves conduct ransomware operations but also against those that provide a safe haven for criminal groups, thereby violating their obligations under international human rights law and the Corfu Channel rule.

Unlike countermeasures, necessity as a circumstance precluding wrongfulness does not require a prior unlawful act. On the grounds of necessity, the wrongfulness of a breach of an international obligation can be precluded where the conduct in violation

[63] African Union (n 61) para 16.

[64] Under this view, the acts of private actors must therefore be attributed to a state. On the content of the right to self-defence in the *jus ad bellum*, see African Union (n 61) para 43; T van Benthem and C Tams, 'Regulating Ransomware Through International Law' (2024) Report of the Scottish Council on Global Affairs 31–32 <https://scga.scot/wp-content/uploads/2024/02/Ransomware-Report-Final-January-2024.pdf> accessed 9 April 2025.

[65] International Law Commission, 'Articles on the Responsibility of States for Internationally Wrongful Acts' (2001) arts 49–54; for further analysis, see Talita Dias, 'Countermeasures in International Law and Their Role in Cyberspace' (2024) Chatham House Research Paper 9–32.

is the only way for the state to safeguard an essential interest against a grave and imminent peril, and it does not seriously impair an essential interest of the state or states towards which the obligation exists, or of the international community as a whole.[66] Because its potential for abuse is significant, this ground must be approached with caution. In this vein of caution and exceptionality, the position of the Netherlands considers that 'the ground of necessity may be invoked only in exceptional cases where not only are there potentially very serious consequences, but there is also an essential interest at stake for the state under threat. What constitutes an "essential interest" is open to interpretation in practice, but in the government's view services such as the electricity grid, water supply and the banking system certainly fall into this category.'[67]

What can be gleaned from this overview is, first, that, as the ransomware threat grows, states may face increasing pressure to undertake offensive extraterritorial cyber operations against ransomware actors, and second, that the legality of their measures would depend on the interpretation of international legal obligations and their exceptions, and circumstances precluding wrongfulness under the law of state responsibility. The contours of both substantive obligations and circumstances precluding wrongfulness remain contested.

States consistently signal their commitment to international law. For instance, Australia has, since its first Cyber Security Strategy in 2016, affirmed that '[a]ny measure used by Australia in deterring and responding to malicious cyber activities would be consistent with our support for the international rules based order and our obligations under international law'.[68] One of the operational principles enshrined in the United Kingdom's 'Responsible Cyber Power in Practice Policy Paper' is that 'operations are conducted in a legal and ethical manner, in line with domestic and international law and our national values'.[69] Commitment to the international legal system necessitates further clarification and agreement on the content of the law, limiting as far as possible grey areas which may come into tension with state sovereignty and foreseeably lead to international escalation. And while grey areas remain, as they do in many fields of national and international law, it bears emphasis that the international law discussion is steadily growing in sophistication in national position papers and multi-stakeholder

[66] International Law Commission, 'Articles on the Responsibility of States for Internationally Wrongful Acts' (2001) art 25.

[67] Government of the Kingdom of the Netherlands, '*Appendix: International Law in Cyberspace*' (2019) 7–8; for further analysis, see Przemysław Roguski, 'Application of International Law to Cyber Operations: A Comparative Analysis of States' Views' (2020) *The Hague Program for Cyber Norms Policy Brief* 20–21.

[68] Australia, *2016 Cyber Security Strategy* 27–28.

[69] UK National Cyber Force, 'Responsible Cyber Power in Practice' (*GOV.UK*, 2023) <https://www.gov.uk/government/publications/responsible-cyber-power-in-practice/responsible-cyber-power-in-practice-html#:~:text=What%20this%20means%20in%20practice,exposing%20hostile%20activity%20and%20wrongdoing> accessed 9 April 2025.

initiatives.[70] A more centralized approach to this clarification would be an important next step. This could serve as a signalled commitment to the international legal system, and a capacity- and confidence-building measure between states and other stakeholders.

5. CONCLUSION

The ransomware ecosystem adapts and evolves, and the threat it poses to societies worldwide continues its upward trajectory. As states continue to debate, both nationally and at the international level, the most effective approaches to counter ransomware criminality, international law must remain a central consideration for both policy-makers and those implementing domestic policies. International law requires states to act in the face of mounting ransomware risks. At the same time, it provides important constraints on state action.

This paper argued that positive obligations under international law compel states to take effective measures to protect individuals under their jurisdiction and other states from ransomware harms, including harms originating in non-state criminal groups. It reviewed measures already undertaken by states that are capable of discharging these positive obligations – the adoption of legislative frameworks for criminalization and reporting, whole-of-society cyber hygiene training and protective measures for critical infrastructure providers, among others. Effectively discharging positive obligations would require a comprehensive approach to protection and continuous updating of domestic measures in light of the evolving ransomware ecosystem.

Beyond measures aimed at domestic resilience-building, states may face the pressure of adopting a more ‘aggressive’ approach to the threat posed by ransomware groups, given their frequent operation from jurisdictions unwilling to take meaningful enforcement action. In crafting any potential extraterritorial measures to interfere with ransomware criminality, states must carefully consider their international law obligations to abstain from unlawful uses of force, coercive interferences and unlawful effects on the sovereignty of other states. Whether a particular activity breaches these negative obligations and whether their breach may be justified would depend on the legal interpretation of the relevant rules, many of which remain pixelated. Especially in areas of heightened geopolitical sensitivity, states must exercise particular caution. One important aspect of being cautious – and responsible – in countering ransomware is to commit to the meaningful clarification of the relevant international law rules.

[70] See eg ‘Tallinn Manual Process’ (CCDCOE) https://ccdcoe.org/research/tallinn-manual/ (accessed 9 April 2025); ‘Oxford Process on International Law Protections in Cyberspace’ (*The Oxford Process*) <https://www.elac.ox.ac.uk/the-oxford-process/> accessed 9 April 2025; ‘International Cyber Law in Practice: Interactive Toolkit’ (*CyberLaw Toolkit, CCDCOE*) <https://cyberlaw.ccdcoe.org/wiki/Main_Page> accessed 9 April 2025.

CyCon 2025: The Next Step
17th International Conference on Cyber Conflict
C. Kwan, N. Gratzer, K. Podiņš, M. Tolppa (Eds.)
2025 © NATO CCDCOE Publications, Tallinn

Permission to make digital or hard copies of this publication for internal use within NATO and for personal or educational use when for non-profit or non-commercial purposes is granted provided that copies bear this notice and a full citation on the first page. Any other reproduction or transmission requires prior written permission by NATO CCDCOE.

Countering Ransomware: Government Responses in a Comparative Perspective

Roxana Radu
Associate Professor of Digital Technologies and Public Policy
Blavatnik School of Government
Hugh Price Fellow, Jesus College
University of Oxford
Oxford, United Kingdom
roxana.radu@bsg.ox.ac.uk

Abstract: Ransomware is currently a top national security threat in many countries around the world. From the disruption of critical infrastructure providers in the US in 2021 to the 2022 paralysis of governmental systems in Costa Rica, ransomware has affected millions of people as direct or indirect victims of extortion practices, data theft, and information access restrictions. Exploring how governments have responded to ransomware since its surge in 2020, this paper expands on current literature that analyzes individual incidents or isolated responses. By incorporating new data from four cases—Australia, Costa Rica, France, and Singapore—this study provides a comprehensive overview of global trends in ransomware mitigation. It introduces an analytical framework based on five levers, ranging from technical to political. The findings underscore a dual focus on improving government coordination through policy centralization and responsibility-sharing while reinforcing public-private partnerships. Across the cases examined, ransomware responses have been multifaceted yet closely aligned with each country's overall cybersecurity posture.*

Keywords: *ransomware, cybercrime, mitigation, national cybersecurity strategies, offensive capabilities*

* This publication arises from research funded by the John Fell Oxford University Press Research Fund (grant no 14333).

1. INTRODUCTION

Since 2020, ransomware has disrupted critical sectors around the world, causing widespread societal and economic harm. The 2021 Colonial Pipeline attack was among the first to make the headlines in a series of hundreds targeting critical infrastructure. A ransomware group called DarkSide was behind this attack on the payment processing system of Colonial Pipeline, resulting in fuel shortages across the eastern United States and impacting millions of businesses and consumers (Kerner 2022; Easterly and Fanning 2023). Two weeks later, Conti's ransomware attack on Ireland's Health Service Executive crippled healthcare services nationwide, jeopardizing patient care and forcing the cancellation of critical medical procedures (Winder 2021). By 2022, ransomware had escalated to the level of national crisis, as seen in Costa Rica, where Conti's attack on government institutions forced the government to declare a "state of emergency," a first for a ransomware attack. Multiple government institutions, including the Ministry of Finance, had their essential services, such as tax collection and customs processing, disrupted for weeks (Murray 2022).

As this brief overview of highly disruptive incidents shows, ransomware attacks have not been confined to specific regions or sectors. They have permeated global systems, including critical supply chains, both physical and virtual. In July 2021, South Africa's Transnet fell victim to a ransomware attack that disrupted port operations, including the Port of Durban. Attackers used strains like "Death Kitty" to encrypt files, grinding logistics to a halt and illustrating the fragility of critical infrastructure (Njini and Viljoen 2021). That same month, a REvil attack on a key United States software vendor, Kaseya, exploited software vulnerabilities to infect more than 1,500 downstream companies. Retailers, manufacturers, and other businesses worldwide faced operational paralysis, including nurseries, schools, pharmacies, and supermarkets in 17 countries, from Sweden to New Zealand, revealing a "new threshold of collective vulnerability" (Radu 2021). Ransomware has firmly established itself as the dominant global cyber threat over the past four years (ENISA 2024), drawing significant international attention and rising on the political and diplomatic agenda. Its prominence grew conspicuously after the June 2021 Biden-Putin summit in Geneva, where it became a key focus of negotiations.

The highly lucrative and adaptable modus operandi of ransomware groups has been driven in large part by the rise of ransomware-as-a-service (RaaS) (Blessing et al. 2022). This model allows cybercriminals to lease advanced ransomware tools and take a cut of the profit, making sophisticated attacks accessible to those with minimal technical expertise. Double extortion—encrypting data while threatening to leak it—has become commonplace in cybercrime, as attackers move to directly blackmailing victims in some cases. RaaS has professionalized the industry, featuring specialized roles like access brokers and distributors within structured networks (NCA and NCSC 2024). These platforms provide customer support and profit-sharing schemes, making ransomware scalable. Despite some operators shutting down (Murray 2022) or being arrested (NCA 2024), over 70 groups (Rapid7 2024; CyberInt 2024) continue to operate from jurisdictions with weak law enforcement cooperation, enabling them to act with impunity.

For these reasons, governments across the globe have faced significant challenges in keeping up with the increasing sophistication and expanding reach of cyberattacks. Despite efforts to combat these threats, cybercriminals are continuously evolving their tactics, swiftly adapting to new security measures in what has often been described as a perpetual game of "whack-a-mole" (NCA and NCSC 2024). This paper examines the public responses to these challenges between 2021 and 2024, offering a structured framework to analyze the levers available to governments. Section 2 delves into the rapid expansion of ransomware attacks, exploring their broader societal impact and the increasing recognition of the harm they inflict on individuals and critical infrastructure. Section 3 explores key policy and academic debates, setting the stage for the analysis by introducing the five levers examined in this study. Section 4 presents the findings, highlighting both commonalities and differences across four different jurisdictions. Finally, the concluding section summarizes the key insights and their broader implications.

2. UNDERSTANDING THE EVOLVING RANSOMWARE THREAT AND ITS HARMS

Despite increased sophistication, ransomware remains largely opportunistic. It relies mostly on "spray and pray" tactics—automated attacks that indiscriminately target numerous systems using common exploits and affecting all systems that lack security measures rather than precisely targeting particular ones. The vast majority of attackers exploit vulnerabilities in unpatched systems or remote access to systems without multi-factor authentication (Rapid7 2024). Only a small percentage resort to zero-day vulnerabilities, or faults not yet known to the manufacturers.[1] Since 2020, over 130 ransomware strains have been identified, with 95 percent of attacks targeting Windows-based systems (VirusTotal 2021). The number of reported active ransomware groups varies, depending on the source. CyberInt (2024) reports an increase from 68 groups in 2023 to 95 in 2024. Rapid7 identifies 75 active groups, including 33 new or rebranded ones. These groups extort their victims through data leaks, resulting in 5,477 posts across leak sites (Rapid7 2024). As of February 2025, Ransomwarelive (2025) documented 238 active ransomware groups.

[1] A notable exception is the Clop group, which intensified its activities in 2023 by exploiting a single zero-day vulnerability that, they claimed, breached 130 organizations (Gatlan 2023).

Many of these groups operate on RaaS platforms, whose developers take a percentage of every successful ransom payment. Average payouts skyrocketed from US$812,380 in 2022 to US$1,542,333 in 2023 (Sophos 2023). Collaboration among ransomware gangs has further enhanced the capabilities of these attacks. LockBit provided a prime example of that in 2023, when it adopted 25 percent of leaked Conti code and released a newly built encryptor to replace its proprietary one (Constantinescu 2023). This cooperative approach, combined with the financial incentives of RaaS, turns ransomware into a criminal activity that keeps pushing boundaries.

Since 2020, the healthcare sector has been particularly vulnerable to ransomware attacks due to its reliance on sensitive data and legacy software. Healthcare systems, in particular, became attractive targets during the COVID-19 pandemic because of their critical nature: one in three health institutions reported at least one ransomware attack in 2020 (Mishra 2024). By 2024, the business services sector became the most targeted, accounting for 24.1 percent of ransomware cases, followed by retail at 15.2 percent and manufacturing at 10.5 percent. A notable shift from 2023 is a 50-percent increase in ransomware incidents within the construction industry, which rose to fourth place, ahead of the financial, education, and healthcare sectors, which had been more heavily targeted in 2023 (CyberInt 2024).

In recent years, the harm caused by ransomware has started to be more clearly understood, though challenges in data availability and underreporting persist. Much of the available data is concentrated in the United States, which skews the broader global picture. Initially, reporting on ransomware focused predominantly on financial losses, such as extortion payments and business interruptions and recovery costs. However, there has been a growing recognition of ransomware's broader societal consequences and cyber harms, including disruptions to daily life and services, as well as erosion of public trust and internal morale (Agrafiotis et al. 2018). In a 2024 UN Security Council briefing, the director-general of the World Health Organization referred to ransomware attacks on hospitals and health facilities as "issues of life and death" (Mishra 2024).

These societal harms are now a focal point in academic and policy discussions on the topic, as researchers and NGOs have started collecting systematic data and exploring the experience of victims (MacColl et al. 2024; Virtual Routes 2025). The CyberPeace Institute (2021) has documented the short- and long-term effects of cyberattacks on healthcare, from the immediate disruptive impact on service and patient care to the less visible impact on the mental health of healthcare professionals and IT specialists. An academic study looking at the impact of the first ransomware incident to make the headlines—the WannaCry attack from 2017—showed a significant decrease in the activity of the hospitals infected across the National Health Service in England.

Over the week of the attack, there were 13,500 appointments cancelled, 1,100 fewer emergency department admissions, and 2,200 fewer elective admissions (Ghafur et al. 2019).

The broader consequences of ransomware extend beyond the immediate disruption of services, particularly within public sectors, where recovering from a ransomware attack also diverts valuable resources from other priorities (MacColl et al. 2022; Martin 2024). While the downtime or interruption post-attack can vary significantly—from an average of 24 days for businesses and organizations in the US (Statista 2024) to months in the case of Costa Rica (Murray and Srivastava 2022)—other effects last for years. Reduced trust in government has been evidenced in the aftermath of a ransomware attack against a Düsseldorf hospital, in particular among segments of the population exposed to the attack (Shandler and Gomez 2022).

However, there is no consistent data collection to allow for a comprehensive analysis. Existing research on the topic has offered fragmented and inconclusive evidence regarding the proactive measures adopted by technologically advanced nations (primarily the United States, the United Kingdom, and the European Union). My contribution addresses this major gap by examining evidence from four jurisdictions—Australia, Costa Rica, France, and Singapore—across four continents. These four countries have various levels of cybersecurity maturity, regulatory stewardship, and resilience. All four have publicly acknowledged the threat that ransomware poses to national security, as a first step in crafting their ransomware responses. Each country offers insights into varying levels of preparedness, legal framework development, and institutional arrangements designed to counter ransomware.

3. HOW HAVE GOVERNMENTS RESPONDED?

Despite abundant policy documents and measures to counter ransomware, research on what has guided the government responses remains sparse. Many case studies of previous ransomware attacks have been used as evidence to prioritize the focus on protecting critical infrastructure, particularly in Australia and the UK (Department of Home Affairs, 2021; UK Government, 2024). The existing scholarly literature primarily identifies general trends and debates, yet it offers limited insight into how these are translated into concrete government actions. This section clarifies what has materialized so far and how these elements inform the identification of relevant levers in government action.

Three key debates on ransomware have structured the policy conversations and continue to underpin many of the policy tools under discussion around the world:

1) the criminalization of ransomware; 2) the role of ransomware insurance; and 3) mandatory reporting requirements. These debates introduce new variables for how to tackle the ransomware threat through legal, economic, and regulatory measures.

A. Criminalization of Ransomware and Crypto Payments

A major debate centers on whether ransomware should be recognized as a distinct criminal offense. This issue gained prominence during negotiations for the recently adopted UN Convention on Cybercrime. Proponents argue that ransomware's unique characteristics within the typology of cyberattacks, such as its extortion-based model and rapid evolution, justify criminalizing it as a specific offense (Robles-Carrillo 2023). Critics, however, caution that such an approach carries practical challenges, given the diverse and constantly evolving forms of ransomware (Robles-Carrillo 2023). In Australia, national discussions on the topic date back to 2021 (Department of Home Affairs 2021). In accordance with the Ransomware Action Plan, the 2024 Cyber Security Bill introduces a stand-alone offense for all forms of cyber extortion and a stand-alone offense for cybercriminals targeting critical infrastructure.

The association with cryptocurrency exchange action has been widely discussed, in an effort to target the financial infrastructure that enables ransomware actors to profit from their attacks. Cryptocurrency exchanges—typically underregulated—facilitate the conversion of illicit crypto ransoms into real-world currency (Alper 2021; TRM 2021). By criminalizing the use of cryptocurrencies in ransomware payments, authorities aim to disrupt the flow of illicit transactions, making it more difficult for cybercriminals to launder money and profit from their activities. This includes measures such as requiring cryptocurrency exchanges to comply with anti-money-laundering regulations, conducting thorough know-your-customer checks, and monitoring suspicious transactions. Such measures are two-fold. On the one hand, they aim to reduce the effectiveness of ransomware campaigns by targeting the financial systems that support them; on the other, they seek to increase the accountability of cryptocurrency platforms in order to prevent their misuse. Targeted action in the area of payment tracing has shown significant progress in 2024 (Chainalysis 2025).

B. Role of Ransomware Insurance

The second debate concerns the role of ransomware insurance as a policy tool to mitigate attacks. Critics argue that it creates perverse incentives by fostering a private market for mitigation and encouraging ransom payments, which embolden cybercriminals (Dudley 2019; Lubin 2022). Insured businesses may also opt to pay ransoms quietly rather than report incidents, complicating law enforcement efforts (Blessing et al. 2022). By contrast, advocates emphasize the benefits of ransomware insurance, particularly for offsetting financial risks faced by large organizations.

Research by Mott et al. (2023) highlights how cyber insurance can act as governance, requiring organizations to meet higher security standards as a condition of coverage and rewarding good risk management. However, challenges persist, including rising loss ratios for insurers and ethical concerns over financing criminal groups (Pauch 2023). O'Connell (2023) advocates banning ransomware payment reimbursements altogether, arguing that this could deter future attacks. In France, this debate has shaped the regulatory approach to allow the insurability of cyber ransoms under the Orientation and Programming Law (2023). However, this is strictly contingent on reporting the incident to authorities within 72 hours, a requirement that strikes a balance between risk mitigation and accountability (Ministère de l'Économie 2023).

C. Mandatory Reporting Requirements

The third debate addresses the issue of underreporting and the limited sharing of information about vulnerabilities, both of which hinder effective policy responses. Mandatory reporting is increasingly viewed as a solution to these challenges. In the EU, the NIS 2 Directive introduces stricter reporting obligations for entities across critical and essential sectors, requiring them to notify national authorities of significant cybersecurity incidents within 24 hours of detection. This directive is a key component of the EU's regulatory stewardship on cybersecurity, aiming to harmonize practices across member states to ensure a higher level of resilience and preparedness. In Australia, the recently enacted Cyber Security Bill mandates reporting of ransomware payments to the Australian Signals Directorate within 72 hours.

D. A New Framework of Analysis

The debates presented above highlight the need to act at the legal and regulatory levels. In addition to these dimensions, implementing technical measures and collaborating internationally to counter ransomware can be important levers for governments to tackle the complex challenge of ransomware. Building on these, the following framework of analysis was developed for this comparative study (see Table I).

This framework is multi-dimensional, designed to encompass a wide array of strategies and policies adopted between January 2021 and September 2024, which are categorized as part of technical, institutional, regulatory, legal, or political levers. By mapping out these strategies, the framework enables a deeper understanding of how governments approach cybersecurity, particularly in the context of countering evolving threats like ransomware. The categorization is grounded in qualitative research, with data collected between April and September 2024 as part of the JFF project conducted at the University of Oxford.

TABLE I: FRAMEWORK OF ANALYSIS BASED ON FIVE LEVERS COVERING DOMESTIC AND INTERNATIONAL ACTION

Lever	Description
Technical	The deployment of advanced technologies and tools to prevent, detect, and recover from cyberattacks, including endpoint protection, intrusion detection, automated threat sharing, and backup solutions.
Institutional	The development and coordination of organizational frameworks, policies, and governance structures to define roles and responsibilities for effective ransomware response and recovery.
Legal	The application of laws and legal instruments to deter, respond to, and mitigate cyberattacks, including criminalizing ransomware, enabling cross-border investigations, and prosecuting ransomware actors operating in different jurisdictions.
Regulatory	The implementation of rules, guidelines, and compliance mechanisms to enforce cybersecurity standards and practices across the public and private sectors, through rules, compliance mechanisms, incident reporting, audits, and adherence to regional frameworks to ensure resilience and preparedness.
Political	The role of political leadership in shaping national and international cybersecurity strategies, allocating resources, and fostering diplomatic efforts for global cooperation against ransomware.

By examining these dimensions, the framework provides valuable insights into the priority areas that governments are addressing, revealing the progress made in key areas such as legislation, institutional development, and international cooperation. Moreover, this comparative analysis reveals where different approaches fall along a spectrum that ranges from defensive to proactive strategies. Finally, this framework serves as a tool for assessing not just the actions taken by individual countries but also the broader trends in governmental responses to cybersecurity challenges.

4. FINDINGS

This section presents the findings of the study, illustrating how the four countries included in the analysis have approached the evolving ransomware threat and discussing their posture in a comparative perspective. In doing so, it advances the scholarship on ransomware, which has primarily focused on individual incidents or isolated responses within a few Western jurisdictions. The new data presented here provides a more comprehensive understanding of global trends in ransomware mitigation, highlighting patterns in public responses to this persistent cybersecurity challenge. Starting from a summary of key developments in each jurisdiction (presented in Table II), I discuss commonalities and differences in ransomware mitigation strategies across the five identified levers. Subsequently, I reflect on the effectiveness of the measures adopted and recent changes in the ransomware ecosystem.

TABLE II: SUMMARY OF KEY DEVELOPMENTS (2021–2024) ACROSS FOUR JURISDICTIONS

Lever	Australia	Costa Rica	France	Singapore
Technical	Pressure testing critical systems Protecting the most valuable datasets Active cyber defense to fight ransomware	Tool for peripheral protection of ministries Periodic analysis of vulnerabilities Cloud computing solutions for the public sector	Focus on domestic industrial capabilities and digital autonomy Separation of defensive and offensive capabilities in combating ransomware	Curated ecosystem of partners for local businesses One-stop ransomware portal Implementing protective DNS Plans to augment ransomware payment tracing capabilities Cybersecurity labeling scheme
Institutional	Executive Cyber Council (public-private threat info sharing) Cyber Incidents Review Board 2024	Cyber Cluster—improvement of cyber ecosystem (2022) Permanent national Security Operations Center (SOC-CR)	CyberCrisis Coordination Centre (since 2018)	Counter Ransomware Task Force (2022) CyberSG TIG Collaboration Centre and the Talent, Innovation and Growth Plan (2023) Government Cyber Security Operations Centre (2022), integrating the Government IT Security Incident Response
Legal	Data Disruption Warrants and Covert Access Obligation 2021 Cyber Security Bill 2024 Intelligence Services and Other Legislation Amendment (Cyber Security) Bill 2024 Security of Critical Infrastructure and Other Legislation Amendment (Enhanced Response and Prevention) Bill 2024	Law 10500—Authorization of interception of cybercrimes • Contingency plans for ICT security in the public sector • Guidelines to reduce the impact and likelihood of ransomware and data extortion incidents in public and private organizations	Guidance and Planning Law of the Ministry of the Interior 2023 Law to secure and regulate the digital space 2024 Transposition of the EU Network and Information Systems Security Act (NIS2) 2024 Regulation on digital operational resilience for the financial sector (DORA) 2022	Online Criminal Harms Act 2023 Cybersecurity (Amendment) Act 2024

Regulatory	Mandatory ransomware payment reporting obligation Ransomware playbook	Strengthening of CSIRT-CR Public entities' obligation to inform CSIRT about incidents	Reporting of incidents (NIS2) Major operational incident reporting obligation (DORA) Critical infrastructure obligations (NIS2) ANSSI can examine compliance with prevention measures	Mandatory cybersecurity Code of Practice for CII operators Licensing of cybersecurity service providers Plans to introduce mandatory obligation to report ransomware payment
Political	Revised National Cyber Security Strategy (2023–2030) Operation Aquila for cybercrime disruption A$9.9 billion committed to boosting AU Signals Directorate's offensive capabilities Co-lead of CRI pillar	National strategy on digital transformation 2023–2027 OAS, EU, CoE cooperation Bilateral agreements in Latin America and beyond CRI member	Stratégie d'Accélération Cybersécurité 2021 Coordination with the EU and NATO Follow-up work as part of the Paris Call (2018) CRI member	Singapore Cybersecurity Strategy 2021 (CSA 2021) ASEAN Voluntary Lead Shepherd on Cybercrime Chairing UN OEWG on Security of and In the Use of ICTs (2021–2025) Operationalization of the ASEAN regional CERT Co-lead of CRI Pillar

A. Commonalities

Across the four cases, a prominent trend is the consolidation of responsibility for ransomware mitigation for critical infrastructure—a departure from previous decentralized efforts in government. By 2024, enhanced horizontal coordination among ministries and public agencies had become essential for a comprehensive approach to ransomware and a more effective deployment of resources and expertise, as indicated in the revised national cybersecurity strategies of the four countries. Streamlining authority through cross-departmental units not only facilitates continuous communication but also supports stronger data-sharing among trusted networks. This aligns with legal mandates for incident reporting and increased protective responsibilities on providers of essential services.

The governments included in this study all recognize the key role of the private sector in safeguarding sensitive data and key services. Consequently, there is a shift towards regulatory measures that mandate the implementation of "security by design" principles across all sectors, thereby complementing and enhancing previously established guidelines. In France, it happens in part through transposing the European NIS2 Directive, whereas in Singapore and Australia, it is supported by legal reforms passed in 2023 and 2024. In addition to amendments to cybersecurity bills, both countries have bolstered the powers of law enforcement and government agencies to ensure more operational tools are available to combat ransomware (Department of Home Affairs 2021; Khan 2024).

Government action has focused not only on proactive defense but also on rapid recovery, to ensure swift organizational rebound following a breach. For example, in 2022, Singapore's Cyber Security Agency released an updated Cybersecurity Code of Practice to aid critical infrastructure owners in countering cyberattacks and enhancing public-private collaboration. Similarly, the Australian Cyber Security Centre offers technical advice and a free Cyber Security Assessment Tool in accordance with the Ransomware Action Plan 2021 (Department of Home Affairs 2021). These efforts are further supported by initiatives aimed at building societal resilience, such as Singapore's centralized ransomware portal and Costa Rica's national cybersecurity education plan.

Finally, all countries included in this study are members of the Counter Ransomware Initiative (CRI), which is currently the world's largest international cyber partnership between governments. Since its launch as a US initiative, the CRI has doubled its membership to over 70 states and refined its governance to enhance resilience, disrupt criminal operations, and shape policy. A key milestone was the 2022 establishment of the International Counter Ransomware Taskforce (ICRTF), which operationalized CRI efforts through intelligence sharing and industry collaboration. By 2023, CRI had evolved into an action-oriented framework—under US coordination—built around three pillars: Policy (co-leads: UK, Singapore), Diplomacy (co-leads: Germany, Nigeria), and ICRTF (co-leads: Australia, Lithuania). The 2023 summit advanced efforts against ransom payments, ransomware infrastructure, and illicit cryptocurrency flows while expanding mentorship for new members, AI-driven countermeasures, and incident response support (Dobell 2024). These actions have positioned the CRI as a credible international framework for developing collective ransomware mitigation strategies.

B. National-Level Variation

At the national level, there is considerable variation in the approaches adopted to counter ransomware, as each of the four countries developed a posture rooted in its own needs and circumstances. While Costa Rica focused extensively on cyber awareness and technical improvements, Australia pursued a disruption-centered direction in both its domestic coordination and international cooperation, particularly as lead of the CRI Disruption Task Force. France and Singapore combined their regional leadership with broader resilience approaches. The different postures and junctures are discussed in more detail below.

The qualitative analysis also reveals divergences along the defensive-offensive spectrum of ransomware responses. Countries with advanced offensive cybersecurity capabilities, like Australia, are more inclined to adopt assertive tactics, proactively disrupting and dismantling cybercriminal networks. By contrast, nations with less mature cybersecurity ecosystems, such as Costa Rica, focus primarily on defensive strategies aimed at enhancing resilience. Their efforts prioritize securing infrastructure, improving incident response mechanisms, and strengthening recovery systems. These variations underscore how national priorities, resource availability, and strategic capacities shape ransomware action.

In Costa Rica, two key public interventions have been prioritized since 2022: technical advancements (including cloud computing solutions for the public sector) and regulatory environment. The country's national strategy on digital transformation (2023–2027) is specifically anchored in the experience of recovering from the Conti attack and presents a comprehensive vision of cybersecurity preparedness. On the technical side, the country has been working on strengthening peripheral protection tools for public authorities. On the regulatory front, there has been a push to improve Costa Rica's ability to access information on cyber incidents and to enable more effective internal reporting. Despite some institutional progress, Costa Rica, like many other Latin American countries, still lacks a full-fledged institutional framework for tackling cybersecurity challenges. On its way to developing one, the nation has pioneered a national cybersecurity education plan.

Australia stands out for its proactive measures in addressing cyber threats, particularly through the establishment of task forces aimed at disrupting cybercriminal networks beyond its borders. The country's commitment to leveraging defensive and active cyber defense capabilities is evident in both its domestic and international approaches. The suite of legal reforms (Parliament of Australia 2024) was foreshadowed in the 2023–2030 Australian Cyber Security Strategy. Aside from the mandatory no-fault, no-liability ransomware reporting obligations, the Cyber Security Bill also enables the government to define mandatory security standards for "connectable products." Under the Security of Critical Infrastructure and Other Legislation (SOCI) Amendment Bill, the government has the power to direct an entity to take action in response to (cyber) incidents. These developments are complemented by technical measures to enhance the preparedness of critical systems and significant investments directed toward the Australian Signals Directorate. The Australian approach is thus both comprehensive and assertive, relying on a strong public-private partnership.

Like Australia, Singapore significantly strengthened its legal framework with the 2024 amendment to its Cybersecurity Act. Key changes include expanding the scope of regulated entities, broadening mandatory incident reporting, and increasing security responsibilities for both virtual and physical systems, including those overseas (CSA 2025). The Act also grants the commissioner of cybersecurity expanded authority to mitigate threats, including directing entities to take or refrain from actions that could reduce risks. Relatedly, the Online Criminal Harms Act from 2023 covers information-sharing and taking action such as blocking access to online content suspected of being used for crime. As a regional cybersecurity leader, Singapore has introduced vetting and certification schemes for cybersecurity and internet-of-things (IoT) products and protective Domain Name Systems (DNS) for government systems, and it is pursuing ransomware payments tracing. Internationally, it leads multiple ransomware initiatives in the Association of Southeast Asian Nations (ASEAN) and plays a key role in the CRI.

In Europe, France has experienced a high number of ransomware attacks between 2021 and 2023. Throughout this time, it has maintained a clear distinction between defensive and offensive capabilities and its tradition of no public attribution. A strong promoter of digital sovereignty, France has focused on domestic industrial capabilities to boost its autonomy, also investing in protections for its governmental systems and talent development locally. This dual approach—bolstering local industry and government cybersecurity—has made public intervention to counter ransomware less of a priority than efforts to advance cyber resilience frameworks. Broader cyber-related obligations and restrictions on businesses were introduced in new laws passed in 2023 and 2024. As a member of the European Union, France has transposed the European directives relevant to cybersecurity (NIS2, DORA) and has been among the first countries to start the horizontal coordination for cyber crises, years before ransomware surged. On the international stage, France has been proactive on advancing cybersecurity in the European Union and has strengthened NATO's cybersecurity cooperation.

This examination of national approaches shows that no single policy lever suffices; instead, a multi-dimensional strategy is essential to combat this evolving threat. From Costa Rica to Australia, the spectrum of proactive cybersecurity measures introduced in recent years has included: 1) expanding the horizontal coordination across government and industry; 2) imposing more obligations on the private sector, particularly critical infrastructure providers; 3) enhancing the powers of public authorities to counter ransomware; and 4) exploring targeted forms of international collaboration (e.g., CRI). These diverse efforts reflect a global recognition that ransomware mitigation necessitates a combination of legal, strategic, and operational responses. But how effective have these levers been?

C. Discussion

Evaluating the effectiveness of ransomware mitigation strategies is challenging in today's cyber ecosystem. While efforts have concentrated largely on reducing vulnerabilities at entry points, an emphasis must also be placed on securing the exit points—specifically, the data exfiltration methods and monetization techniques employed by attackers. According to Chainalysis (2025), the notable drop in ransom payments in 2024 can be attributed to intensified efforts against the money-laundering infrastructure, coupled with more advanced defenses and improved response plans implemented by governments.

Despite variations in sources, the available data suggests a decline in successful ransomware attacks in three of the jurisdictions analyzed. In Australia, incidents decreased slightly, from 107 in 2023 to 101 in 2024 (CyberInt 2023, 2024). France reported 130 incidents, a 21-percent reduction from the previous year (CyberInt 2024). Singapore's numbers remained stable, with 132 incidents recorded in both 2022 and 2023 (SPOR 2024), although 2024 data is not yet available. For Costa Rica, data is also missing; however, following the Conti attack on government services in 2022, the country continued to be the second most affected country in Central America, experiencing over 5,000 attempted attacks in 2023 (Kaspersky 2023). In 2024, a new wave of ransomware incidents targeted key institutions in the country (Tico Times 2024).

Incident response data—albeit only partially available and unevenly distributed—indicates important shifts in the ransomware ecosystem. Coveware's latest quarterly report (2025) indicates that a significantly smaller proportion of the victims are paying ransoms: one-quarter of the affected companies, an all-time low. Moreover, the median payment amounts are decreasing. The tracking of ransomware payments in cryptocurrency reveals a 35-percent decline, from US$1.25 billion in 2023 to US$813 million in 2024 (Chainalysis 2025). This change is attributed to the diminished operational capability and market reputation of prominent RaaS groups targeted by coordinated law enforcement operations in 2024.

However, the overall threat persists as new actors have stepped in (Symantec 2025; Coveware 2025). This study shows that governments are also adapting, through the consolidation of public sector responsibilities and improved data-sharing mechanisms, prioritizing threat intelligence and cross-sector partnerships. The shift is grounded in a broader effort to bolster cyber resilience, by clarifying legal obligations, streamlining institutional powers, and reinforcing critical infrastructure preparedness. A "whole-of-society" resilience approach is starting to take shape through the implementation of talent development programs and cybersecurity skills initiatives.

5. CONCLUSION

This study introduced a novel, multi-dimensional framework to analyze government responses to ransomware, incorporating recent qualitative data (2021–2024) from four jurisdictions. By examining progress in technical measures, legislation, regulation, institutional development, and international collaboration, the analysis reveals convergence around key action areas (critical infrastructure protection; security by design approaches) and variation according to the level of maturity and cyber posture of each jurisdiction. While national priorities and resources vary, Australia, Costa Rica, France, and Singapore share an emphasis on both strengthening internal government coordination and enhancing government-industry partnerships in the fight against ransomware.

The analysis reveals a growing centralization of government responsibilities, driven by a wider cyber resilience impetus. New regulatory measures, such as mandatory incident reporting and enhanced data-sharing requirements, are reshaping the partnership between governments and industries. In the face of this persistent threat, both public authorities and industry are adopting more mature and increasingly strategic responses. However, countries differ significantly when it comes to their priorities and alignment with national posture and circumstances, which range from cyber awareness to deploying offensive capabilities to disrupt ransomware networks. While some countries with advanced cyber capabilities favor proactive disruption, others prioritize defensive resilience. Yet the effectiveness of individual measures remains difficult to ascertain due to the lack of harmonized data.

In the future, more attention needs to be directed towards evaluating the mitigation efforts at the national level, through data collection and systematic policy impact assessments. Policy-makers should conduct comprehensive evaluations of ransomware-targeting measures to gauge their success and identify unintended consequences. Governments can learn from one another by analyzing the incentive structures they establish, but the wide variety of mitigation measures warrants more systematic comparative analyses at the regional level. Finally, there is a pressing need for academic research to broaden the perspective by providing deeper qualitative insights and evidence-based analysis.

REFERENCES

Alper, Alexandra. 2021. "Biden Sanctions Cryptocurrency Exchange over Ransomware Attacks." *Reuters*, 21 September. https://www.reuters.com/business/finance/biden-sanctions-cryptocurrency-exchange-over-ransomware-attacks-2021-09-21/.

Blessing, Jenny, Jules Drean, and Sarah Radway. 2022. “Survey and Analysis of US Policies to Address Ransomware.” *MIT Science Policy Review*.

Chainalysis. *2025 Crypto Crime Report*. 5 February. https://www.chainalysis.com/blog/crypto-crime-ransomware-victim-extortion-2025/.

Constantinescu, Vlad. 2023. “Lockbit Ransomware Gang Switches to Conti-Based Encryptor.” *BitDefender News*, 3 February. https://www.chainalysis.com/blog/crypto-crime-ransomware-victim-extortion-2025/.

Coveware. 2025. “Will Law Enforcement Success against Ransomware Continue in 2025?” *Quarterly Report*, 4 February. https://www.coveware.com/blog/2025/1/31/q4-report.

CSA. 2021. *The Singapore Cybersecurity Strategy 2021*. Government of Singapore.

CSA. 2024. *Singapore Cyber Landscape 2023*. 30 July. https://cyberint.com/blog/research/ransomware-annual-report-2024/.

CyberInt. 2025. *Ransomware Annual Report 2024*. 13 January. https://cyberint.com/blog/research/ransomware-annual-report-2024/.

CyberInt. 2023. *Ransomware Trends Report 2023*. 7 April. https://cyberint.com/blog/research/ransomware-trends-and-statistics-2023-report/.

CyberPeace Institute. 2021. *Playing with Lives: Cyberattacks on Healthcare Are Attacks on People*. CyberPeace Institute.

Department of Home Affairs. 2021. *Ransomware Action Plan*. Australian Government. https://www.homeaffairs.gov.au/about-us/our-portfolios/cyber-%20security/strategy/australias-ransomware-action-plan.

Dobell, Adam. 2024. “The International Counter Ransomware Initiative: From Forming and Norming to Performing.” *Center for Cybersecurity Policy and Law*, 24 September.

Dudley, Renee. 2019. “The Extortion Economy: How Insurance Companies Are Fueling a Rise in Ransomware Attacks.” *ProPublica*, 27 August.

ENISA. 2024. *Threat Landscape Report 2024: June 2023–June 2024*. September. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024.

Gatlan, Sergiu. 2023. “Clop Ransomware Claims It Breached 130 Orgs Using GoAnywhere Zero-Day.” *Bleeping Computer*, 10 February. https://www.bleepingcomputer.com/news/security/clop-ransomware-claims-it-breached-130-orgs-using-goanywhere-zero-day/.

Ghafur, S., S. Kristensen, K. Honeyford, G. Martin, A. Darzi, and P. Aylin. 2019. “A Retrospective Impact Analysis of the WannaCry Cyberattack on the NHS.” *npj Digital Medicine* 2(98). https:/doi.org/10.1038/s41746-019-0161-6.

Khan, Azfer A. 2024. “Reconceptualizing Policing for Cybercrime: Perspectives from Singapore.” *Laws* 13(4): 44. https://doi.org/10.3390/laws13040044.

Lubin, Asaf. 2022. “The Law and Politics of Ransomware.” *Vanderbilt Journal of Transnational Law* 55: 1177.

MacColl, Jamie, Pia Hüsch, and Jason R. C. Nurse. 2022. *Beyond the Bottom Line: The Societal Impact of Ransomware*. RUSI, 14 November. https://www.rusi.org/explore-%20our-research/publications/commentary/beyond-bottom-line-societal-impact-%20ransomware.

Martin, Ciaran. 2024. “On the Matter of the British Library.” 24 January. https://ciaranmartin.substack.com/p/on-the-matter-of-the-british-library.

Ministère de l'Économie, des Finances et de l'Industrie (Ministère de l'Économie). 2021. *Stratégie d'accélération cybersécurité*. Le Gouvernement de la République Française. https://www.entreprises.gouv.fr/fr/strategies-d-acceleration/strategie-d-acceleration-cybersecurite.

Ministère de l'Économie, des Finances et de l'Industrie (Ministère de l'Économie). 2023. "Lettre de la DAJ – La loi d'orientation et de programmation du ministère de l'Intérieur." https://www.economie.gouv.fr/daj/lettre-de-la-daj-la-loi-dorientation-et-de-programmation-du-ministere-de-linterieur.

Mishra, Vibhu. 2024. "Cyberattacks on Healthcare: A Global Threat That Can't Be Ignored." *UN New*s, 8 November. https://news.un.org/en/story/2024/11/1156751.

Mott, G., S. Turner, J. R. Nurse, J. MacColl, J. Sullivan, A. Cartwright, and E. Cartwright. 2023. "Between a Rock and a Hard(ening) Place: Cyber Insurance in the Ransomware Era." *Computers & Security* 128: 103162.

Murray, Christine, and Mehul Srivastava. 2022. "How Conti Ransomware Group Crippled Costa Rica—Then Fell Apart." *Financial Times*, 9 July. https://www.ft.com/content/9895f997-5941-445c-9572-9cef66d130f5.

NCA. 2024. "International Investigation Disrupts the World's Most Harmful Cyber Crime Group.". https://www.nationalcrimeagency.gov.uk/news/nca-leads-international-investigation-targeting-worlds-most-harmful-ransomware-group.

NCA and NCSC. 2024. *Ransomware, Extortion and the Cyber Crime Ecosystem*. White Paper. https://www.ncsc.gov.uk/files/White-paper-Ransomware-extortion-and-the-cyber-crime-ecosystem.pdf.

Njini, Felix, and John Viljoen. 2021. "Transnet Declares Force Majeure at SA Ports over Cyberattack." 27 July. https://www.news24.com/Fin24/transnet-declares-force-%20majeure-at-sa-ports-over-cyber-attack-20210727.

O'Connell, Sean. 2023. "To Ban Ransomware Payments or Not to Ban Ransomware Payments: The Problems of Drafting Legislation in Response to Ransomware." *Journal of International Business & Law* 22: 151.

Parliament of Australia. 2024. *Cyber Security Legislative Package*. https:\www.aph.gov.au\Parliamentary_Business\Committees\Joint\Intelligence_and_Security\CyberSecurityPackage.

Pauch, Dariusz. 2023. "Ransomware Attacks as a Cybersecurity Insurance Coverage Threat." *Humanities and Social Sciences* 30(2): 99–107.

Ransomware Task Force. 2021. "Combating Ransomware." Institute for Security and Technology. https://securityandtechnology.org/ransomwaretaskforce/report/.

Rapid7. 2024. *Threat Landscape Statistics: Ransomware Activity, Vulnerability Exploits, and Attack Trends*. Rapid7.

Robles-Carrillo, Margarita A., and Pedro García-Teodoro. 2022. "Ransomware: An Interdisciplinary Technical and Legal Approach." *Security and Communication Networks* 2022 (1): 2806605.

Shandler, Ryana, and Miguel Alberto Gomez. 2022. "The Hidden Threat of Cyber-Attacks—Undermining Public Confidence in Government." *Journal of Information Technology & Politics* 20(4): 359–74.

Sophos. 2023. *Ransomware Payouts and Recovery Costs Went Way Up in 2023*. Report.

SPOR. 2024. "Cybersecurity and Digital Resilience." 8 November. https://spor.performancereports.gov.sg/businesses/strong-and-resilient-economy/data-and-cyber-security.

Statista. 2024. "Average Duration of Downtime after a Ransomware Attack at Organizations in the United States from 1st Quarter 2020 to 2nd Quarter 2022." https://www.statista.com/statistics/1275029/length-of-downtime-after-ransomware-%20attack-us.

Symantec. 2024. “Ransomware 2025: Attacks Keep Rising as Threat Shows. Its Resilience.” 20 February. https://www.security.com/threat-intelligence/ransomware-trends-2025.

TRM. 2021. “OFAC Takes First Action against Cryptocurrency Exchange and Issues Updated Ransomware Advisory.” https://www.trmlabs.com/post/ofac-takes-first-action-%20against-cryptocurrency-exchange-and-issues-updated-ransomware-advisory.

VirusTotal. 2021. “Ransomware in a Global Context.” VirusTotal.

Winder, Davey. 2021. “The Five Most Important Ransomware Attacks of 2021.” *Raconteur*. https://www.raconteur.net/technology/the-five-most-important-ransomware-attacks-of-2021.

CyCon 2025: The Next Step
17th International Conference on Cyber Conflict
C. Kwan, N. Gratzer, K. Podiņš, M. Tolppa (Eds.)
2025 © NATO CCDCOE Publications, Tallinn

Permission to make digital or hard copies of this publication for internal use within NATO and for personal or educational use when for non-profit or non-commercial purposes is granted provided that copies bear this notice and a full citation on the first page. Any other reproduction or transmission requires prior written permission by NATO CCDCOE.

The Role of Big Tech Corporations in Military Defence and Civilian Protection: A Case Study of the Russo-Ukrainian War

Clara Cotroneo
Faculty of Governance and Global Affairs, Institute of Security and Global Affairs, University of Leiden,
The Netherlands

Prof. Sarah Leonard
Faculty of Life Sciences and Education, Criminology, Policing and Security Research and Innovation Group,
University of South Wales

Abstract: This paper examines the role of Big Tech corporations in international security – more specifically, in military defence and civilian protection. It investigates the role of Big Tech corporations in conflicts by examining one case: the ongoing conflict between Russia and Ukraine. Researchers in international security and international relations have recently started considering how multinational corporations impact global power balances, international governance and security. Governments and international governmental organizations such as the North Atlantic Treaty Organization (NATO) systematically and strategically cooperate with Big Techs to increase their technology and innovation capacities, capabilities and effectiveness in defence. With technological sovereignty and the adoption of state-of-the-art technologies providing political, economic and military advantages, states and coalitions of governments have boosted their cooperation with the Big Tech industry and have increasingly invested in research and development programmes. Against this backdrop, this paper identifies and examines the main areas where Big Tech corporations have contributed to Ukrainian cyber resilience and civilian protection, using data collected and examined through a content analysis of academic and grey literature and the qualitative analysis of data gathered through four semi-structured interviews. The paper argues that there are six main areas where Big Tech corporations have contributed to Ukrainian cyber resilience and civilian protection: (1) providing cyber threat intelligence, analysis and advice; (2) offering technical support and cyber security solutions; (3) providing assistance and backup when critical or digital infrastructure is disrupted; (4) countering cyber espionage; (5) countering disinformation; and (6) protecting civilians' physical safety and contributing to humanitarian assistance.

Keywords: *cyber resilience, Big Tech corporations, public-private cooperation, military defence, civilian defence, civilian protection*

1. INTRODUCTION

Frequent reports of cyber operations during the ongoing Russo-Ukrainian War have fuelled debates over the cyber dimension of contemporary warfare. Scholarly research on conceptualizations and practices of 'cyber war' has seen sharp disagreements over the meaning and significance of cyber operations in today's warfare theories and practices (Rid 2012; Kello 2013; Delerue 2020; Jacobsen 2021; Neuman 2021). Nevertheless, it is widely agreed that there has been an increase in the use of information and communication technology in warfare, espionage operations, misinformation campaigns and operations interfering with democratic processes. In this context, state actors have intensified their cooperation with the private sector to increase their cyber resilience, cyber defence and deterrence capabilities. Recent research has suggested that Big Tech companies – that is, the most influential technological companies, including Alphabet, Amazon, Apple, Meta and Microsoft – have been supporting state defence and deterrence efforts, including during the conflict in Ukraine (van Benthem 2023). However, research on this topic remains limited. This is primarily because the study of international relations has traditionally been focused on the role of states. Nevertheless, the centrality of state actors on the international stage has increasingly been questioned (Geppert and Dörrenbächer 2014; Babic, Fichtner and Heemskerk 2017). As part of this trend, some researchers have considered the role of multinational corporations in international politics, including, more recently, during the Russo-Ukrainian War. This paper contributes to these debates by focusing on an issue largely overlooked to date: the specific contribution of Big Tech companies to state-level cyber resilience and civilian protection using the case of the Russo-Ukrainian War.

This paper examines two important dimensions of the involvement of Big Tech corporations in the ongoing conflict in Ukraine. The first concerns the Ukrainian government's cooperation with Big Techs to improve its cyber resilience and deterrence capabilities. More precisely, the paper investigates how Big Tech corporations have supported the Ukrainian government against cyber threats emanating from Russia, including cyber attacks against its critical infrastructure, cyber espionage and disinformation campaigns. The second dimension focuses on the initiatives of Big Techs aiming to enhance the protection of and support to Ukrainian civilians. The objective of this paper is twofold. First, it seeks to provide government officials with a mapping of areas for cooperation with Big Techs to enhance state-level cyber resilience. Second, it seeks to provide scholars with a basis for future research on the influence of Big Techs on state security, autonomy and independence. By providing an extensive and detailed mapping of Big Techs' areas of involvement in state-level cyber defence, the paper contributes to practitioners' and scholars' further understanding of Big Techs' capacities, capabilities, power and influence in national and international security.

Context, Rationale and Objectives of the Case Study

Russian military and cyber hostilities against Ukraine date back to at least 2014 – a year that saw the Maidan Revolution, Russia's annexation of Crimea and the conflict in Donbas (Robinson, Jones and Janicke 2015). Having dealt with Russian offensive military and cyber operations for many years, at the time of the full-scale Russian invasion in 2022, the Ukrainian government expected an escalation of hostilities, including via traditional, kinetic military attacks, as well as in and through cyberspace. As part of their strategy to prevent or mitigate the impacts of Russia's cyber attacks, the Ukrainian authorities sought cooperation with Big Tech corporations. Some commentators cited this as an example of agility, resilience and readiness in cyber defence (Lewis 2022; Schulze and Kerttunen 2023). A noteworthy example was the decision of the Ukrainian government to transfer critical and sensitive government data onto the Amazon Web Services (AWS) cloud on the eve of Russia's full-scale invasion. This enabled the protection of critical information and citizen services from the consequences of the military invasion and cyber attacks, thereby enhancing Ukraine's cyber preparedness and ability to execute government functions, decision-making and essential government services.

Russia's cyber operations against Ukraine have intensified as the war has progressed, posing new challenges for the Ukrainian authorities. The CyberPeace Institute recorded almost 3,255 cyber attacks and malicious cyber operations against Ukraine between January 2022 and December 2023. The Computer Emergency Response Team of Ukraine has disclosed that Russia-sponsored cyber attacks have targeted critical Ukrainian infrastructure in the energy, heating and water sectors across 10 regions of the country, disrupting the country's ability to provide essential goods and services to its citizens at home and abroad (CERT-UA 2024). The 2022 Microsoft Intelligence Report reported Russian network intrusion attempts against 128 organizations in 42 countries that maintain alliances with Ukraine (Microsoft 2022b). The main targets have been governments of NATO member states. However, the attacks have also targeted think tanks, humanitarian organizations, and information-technology companies. While the US has been the primary target, Poland, which has largely coordinated the logistical delivery of military and humanitarian aid to Ukraine, has also been a top target, together with the Baltic countries (CERT-EU 2023). According to the same report, 29 per cent of these attacks were successful, among which a quarter led to data thefts. Furthermore, 66 per cent of organizations have modified their cyber security strategies directly in response to the ongoing conflict between Russia and Ukraine (Venafi 2022). From a military perspective, a prime example of the impact of IT service disruption has been the deterioration in the functioning of the Starlink internet service, a pivotal asset for Ukraine's military, following the 2022 invasion (Mozur and Satariano 2024). In strategic terms, Russian cyber attacks and operations have often been executed before or simultaneously with kinetic attacks to amplify their consequences (Geers 2015; Andrew and Geers 2015; Willett 2022). Thus, given the significance of cyber operations during the Russia-Ukraine conflict (Andrew and Geers 2015; Rõigas 2018), this is a particularly relevant case for examining the role of Big Techs in international security, including cyber defence and civilian protection.

2. ANALYTICAL FRAMEWORK AND METHODOLOGY

The analytical framework (Table I) has been designed based on the research questions therein. It draws on the work of Lucas Kello on cyber security and the international order (Kello 2017), the theoretical and methodological framework developed by Babic, Fichtner and Heemskerk (2017), and the theoretical perspective recently advanced by Abels (2024). Data have been collected and analysed through (1) the content analysis of academic and grey literature, including governmental and non-governmental reports (e.g. industry reports), databases and webpages of relevant organizations, and (2) the qualitative analysis of data gathered through four semi-structured interviews with officials from NATO and the European External Action Service, as well as experts.[1]

[1] Interviewees have been selected based on their experience with cyber defence and/or with engaging the private sector in international security policies. Ukrainian public officials and representatives of Big Techs have been contacted. Due to difficulties with engaging these two stakeholder categories, the authors have engaged in extensive desk research, document review and thorough data triangulation, to minimize gaps.

TABLE I. ANALYTICAL FRAMEWORK

Dimension(s) Examined and Research Question(s)	Indicator(s)
Cyber security, cyber resilience and cyber defence	
• How have Big Tech companies contributed to Ukraine's cyber resilience, security and defence? • What have been the main features and benefits of the Big Techs' *modus operandi*? • What has been the additional value of the involvement of Big Tech companies?	Indicator 1. Big Tech companies have provided warnings and/or solutions necessary to shield from or mitigate the impacts of cyber attacks against Ukraine's digital and critical infrastructure Indicator 2. The Big Tech industry has provided support that complemented or substituted national capabilities Indicator 3. Big Techs have offered complementary protective and mitigating solutions, in some cases uniquely available to them
Disinformation and cyber espionage	
• How have Big Tech companies contributed to countering disinformation and cyber espionage? • Have Big Tech companies provided additional or complementary support to the Ukrainian authorities?	Indicator 1. Big Tech companies have put in place measures to counter disinformation Indicator 2. The measures put in place by Big Tech companies could only have been put in place by these companies as owners of media and social platforms
Civilian protection	
• How have Big Tech companies supported civilian protection? • How have Big Tech companies enhanced civilian protection, including the provision of essential services and safe places to civilians in cases of conflict?	Indicator 1. Big Tech companies have enhanced the protection of critical civilian infrastructure Indicator 2. The Big Tech industry has given additional support to government and non- government bodies providing essential services and physical protection Indicator 3. Big Tech companies have contributed to the physical safety of civilians

3. MAPPING AND ANALYSIS OF THE BIG TECH INDUSTRY'S SUPPORT FOR UKRAINIAN CYBER RESILIENCE AND CIVILIAN PROTECTION

This section presents the findings of our case study analysis. Six main areas where Big Tech corporations have contributed to Ukrainian cyber resilience and civilian protection have been identified: (1) providing cyber threat intelligence, analysis and advice; (2) offering technical support and cyber security solutions; (3) providing assistance and backup when critical or digital infrastructure is disrupted; (4) countering cyber espionage; (5) countering disinformation; and (6) protecting civilians' physical safety and contributing to humanitarian assistance.

A. Providing Cyber Threat Intelligence, Analysis and Advice

Immediately before and during the first six months of Russia's full-scale military invasion of Ukraine, the Ukrainian authorities and cyber security agencies were confronted with an unprecedented number of cyber attacks, potentially above their cyber defence capabilities (Beecroft 2022). Under enormous pressure to defend their infrastructure, the Ukrainian authorities requested the assistance of governments sympathetic to their cause to enhance cyber resilience capabilities (Beecroft 2022). In addition to foreign governments, Big Tech corporations have provided support to the Ukrainian authorities to address the effects of Russia-driven cyber attacks.[2] This cooperative strategy between the Ukrainian authorities and Big Tech corporations has been referred to as a 'new form of collective defence', emphasizing the alignment between private sector Big Techs and Ukrainian war efforts, as well as their contributions to Ukrainian cyber defence capacities (Microsoft 2022b). The Big Tech industry has boosted Ukrainian cyber resilience through two types of services: the provision of threat intelligence, analysis and advice and the provision of technical cyber security solutions. Cyber threat intelligence and analysis play a crucial role in enabling governments to understand and navigate the cyber threat landscape. Although risk assessment exercises can be undertaken by national agencies, Big Tech companies are particularly well-placed to conduct such exercises, for two main reasons (Qamar, Anwar and Afzal 2023). First, since they own, manage and operate data centres, Big Tech corporations are equipped with dedicated teams of engineers and technicians who monitor threats to these data centres. Second, tailored threat analysis reports are part of their product portfolio.

To illustrate the cooperation between Big Tech corporations and the Ukrainian government, it is worth noting that, several hours before the Russian launch of missiles and the military invasion of Ukraine in February 2022, Microsoft warned the Ukrainian authorities of a new malware threat (Microsoft 2022a). The company's Threat Intelligence Centre provided the Ukrainian authorities with a tailored cyber threat analysis, which identified potential offensive and destructive cyber attacks targeting Ukraine's digital infrastructure. More recently, the company has also provided threat intelligence to Ukrainian authorities on attacks that were to target Ukrainian military establishments and government agencies.

[2] Interview with author.

This example highlights two important elements of Big Tech corporations' engagement with the Ukrainian authorities. The first element is complementarity. In the context of the conflict, the Ukrainian government needed additional resources for threat identification and assessment, which were provided by foreign governments and Big Tech companies. The added value of Big Techs was specifically to provide tailored cyber intelligence and solutions, which were needed to support or relieve the burden on Ukrainian security agencies. The second element concerns the dynamics of the cooperation and modus operandi of Big Tech corporations. In this example, Microsoft proactively scanned the threat landscape to identify imminent threats against Ukraine's digital and critical infrastructure, in order to inform and propose a solution to the Ukrainian authorities. Furthermore, the identification of cyber threats has been timely, which is important to prevent severe disruptions to the governance and functions of Ukrainian authorities and critical infrastructure.

B. Offering Technical Support and Cyber Security Solutions

To shield national infrastructure from cyber attacks, Ukraine's national authorities also require technical solutions, that is, trusted commercial products (i.e. hardware and software) to protect digital assets from cyber threats, mitigate the propagation and severity of impacts, provide backup and ensure continued functioning in case of disruption to critical and digital infrastructure.[3] The provision of technical support and cyber security solutions is one activity providing revenue to Big Tech companies. The example provided in the previous section of this paper focused on Microsoft's early warning of cyber attacks threatening Ukrainian cyber security resilience. Less than three hours after the threat discovery, Microsoft proposed to the Ukrainian government a technical solution that could be quickly added to the Microsoft Defender anti-malware service and that would allow the Ukrainian government to detect and proactively defend its digital infrastructure against this new threat. The quick exchange and signature of a commercial agreement between the corporation and the Ukrainian government showcases proactiveness, timeliness and the tech companies' ability to provide preventive technical solutions, enhance national cyber resilience and be integrated into government decision-making processes.

The Ukrainian authorities have continued to rely on Big Tech technical solutions for preventing cyber intrusions and cyber attacks following Russia's military invasion. Facing increasing concerns of cyber espionage and cyber intrusions, the Ukrainian authorities signed an agreement with Google regarding the acquisition for their public officials of 5,000 security keys (Ministry of Digital Transformation of Ukraine 2024), a tool that replaces passwords with physical verification. On this occasion, the Ukrainian minister of digital transformation emphasized the 'many years of experience of cooperation with Google' and the support provided by the company to the ministry, which ensured 'the protection of digital infrastructure… crucial in the

[3] Interview with author.

context of a full-scale war' (Ministry of Digital Transformation in Ukraine 2024). Furthermore, the cooperation between the Ministry of Digital Transformation and Google in the area of prevention has included capacity-building initiatives, such as Google-run training courses in cyber security and artificial intelligence for ministry officials. The minister's commentary on the cooperation with Google reveals that Ukrainian public authorities have considered Big Tech corporations to be significant actors in supporting their cyber resilience efforts. The last example is Project Shield (Google 2024), which was created by Google Cloud to provide protection specifically against distributed denial-of-service (DDoS) cyber attacks. This is a preventive tool designed to protect news, elections and human rights websites from DDoS, which would make content unavailable to users. Although the initiative was launched in 2016, Google has extended eligibility for Project Shield to cover various Ukrainian websites, including those of the Ministry of Foreign Affairs and Ministry of Internal Affairs (Huntley 2022), as well as services like Livemap, which helps the civilian population find essential information. About 150 websites in Ukraine, including many news organizations, have been using this service. Big Tech corporations have therefore assisted the Ukrainian authorities and their cyber resilience and cyber defence efforts by providing both responsive and preventive tools in a timely manner.

C. Providing Assistance and Backup When Critical or Digital Infrastructure Is Disrupted

Since at least 2014, Ukraine has faced persistent cyber attacks against critical infrastructure providing essential goods and services. A highly visible case was the attack against the Ukrainian power grids attributed to the Russia-sponsored group Sandworm. This attack received attention worldwide because of the significant power outage that followed, which prompted more serious discussions on the vulnerability of critical infrastructure to cyber attacks, the material impact of such attacks beyond national borders and their significance for international security (Aviv and Ferri 2023; Maschmeyer 2021). Before and during the conflict, tech companies have supported Ukrainian energy companies, including DTEK and Urenergo, to ensure the continued functioning of energy infrastructure (Microsoft 2023). Cloud technologies, such as those provided by Microsoft Azure, have supported the functioning of critical energy infrastructure by allowing centralized data security and accessibility, as well as supporting the continuation of business operations in the Ukrainian banking sector (Microsoft 2022c).

These examples suggest that Big Tech corporations may play an important role in national and international security, thanks to their role in ensuring the continued functioning of national critical infrastructure. In this specific case, they have provided cyber security with preventive and continuity solutions to protect Ukrainian infrastructure, which has mitigated the effects of the attacks emanating from Russia.

In turn, the increased protection of Ukrainian infrastructure has mitigated cascading effects on other critical infrastructure sectors, in Ukraine and abroad. Finally, preventive and continuity measures offload the work of national authorities.[4] Importantly, the role of tech in defence and security has been recognized by foreign governments, as evidenced by the Cybersecurity for Critical Infrastructure Activity project, funded by the US Agency for International Development (USAID 2022). This body fosters cooperation between the private sector and governments to increase the resilience of critical infrastructure. Within the framework of this programme, activities are funded specifically to provide the Ukrainian government with cutting-edge solutions for critical infrastructure protection.

D. Countering Cyber Espionage

Another way in which Big Tech corporations have addressed cyber threats has been the use of civil litigation against malicious cyber actors allegedly responsible for cyber espionage against Ukraine and its political allies. Russian cyber operations have involved cyber attacks, disinformation campaigns and cyber espionage; the latter was intended to steal intelligence relevant to gain military advantages (Štrucl 2022). Actors in cyber espionage against Ukraine include state bodies and state-sponsored groups. The European Repository of Cyber Incidents (EuRepoC) has identified the cyber group APT28 as one of the primary cyber threat actors, allegedly linked to the Chief Intelligence Directorate of the Russian General Staff (GRU). Based on EuRepoC's analytical profiling of APT28, gathering intelligence useful to target national critical infrastructure in the United States, Europe and countries politically aligned with Ukraine has been one of the main activities of the group (EuRepoC 2024). More specifically, this group has focused on information operations and cyber espionage campaigns directed mainly against Ukraine and state entities politically allied with Ukraine, including EU and NATO members. The group's attacks have had direct and indirect impacts, particularly on critical infrastructure and political systems in these states.

As a technical measure, Microsoft took down servers and websites used by this group in 2016 and 2018, executing a court order to disrupt and take control of six internet domains that it had created. In addition, Microsoft initiated legal proceedings against the group in 2016 on the grounds that they had committed an 'internet-based cyber-theft operation' to transfer the group domain names onto Microsoft servers (Schwartz 2017). Thus, this initiative highlights another layer of the interplay between tech corporations and the Ukrainian authorities. In addition to providing cyber security solutions and taking technical measures to disrupt criminal internet infrastructure, Big Tech companies have reverted to formal, traditional instruments to counter cyber opponents, namely civil litigation. The civil litigation case initiated by Microsoft

[4] Interview with author.

shows that tech corporations also use non-technical tools to address state-sponsored and state-targeted malicious cyber activities.

E. Countering Disinformation

Disinformation campaigns have been used to manipulate narratives and polarize the public and political opinion in Ukraine, Russia and other countries, with the aim of destabilizing Ukraine and weakening support for the country. Russia-driven disinformation campaigns have taken place via different types of media, including television and social media platforms. This has had varying degrees of success amongst Ukrainians with partisan and ethnolinguistic ties to Russia but less success amongst other categories of Ukrainian citizens (Erlich and Garner 2023; Golovchenko et al. 2018; Lange-Ionatamishvili et al. 2015). Ukrainian authorities have directly called out social media platforms on their responsibility to counter disinformation. For instance, Ukraine's minister of digital transformation asked the CEO of YouTube to block Russian disinformation campaigns depicting Ukrainians as Nazis and drug addicts (Cerulus 2022).

In practice, Big Tech companies like Apple, Microsoft and Google have removed RT and Sputnik News from their app stores (Dave 2022). Google has restricted the presence of RT and Sputnik in the European Union. Similarly, social media platforms TikTok and Facebook have blocked RT and Sputnik News across Europe (Culliford 2022). Among the technical measures that media companies have put in place to counter disinformation, a relevant example is X's labelling strategy, which involved the labelling of messages containing links to Russian state-affiliated media. X's labelling strategy alerted users to be cautious when they saw, opened or consulted websites that might contain untruthful information. Similarly, Facebook and Instagram have taken measures to globally demote posts with links to Russian state-controlled media, with the same objective of shifting users' attention away from these contents (Culliford and Dang 2022). Furthermore, X has adopted a mirroring strategy by incorporating a fact-checking tool to flag misleading tweets.

The objectives pursued by the Ukrainian authorities through cooperation with tech companies go further than countering Russian disinformation. They have also called for restricting access to social media in Russia in order to generate discontent among young people towards the Kremlin. This was why the Ukrainian minister of digital transformation asked the CEOs of Apple, Google, YouTube and Netflix to restrict or block their services in Russia. The decision of these companies to limit some of their services may indicate that they are taking a clear position in the conflict; the reasons for this require further investigation.

F. Protecting Civilians' Physical Safety and Contributing to Humanitarian Assistance

Big Tech companies have supported the work of international governmental agencies and non-governmental organizations (NGOs) establishing safer environments for Ukrainian civilians. Additionally, they have put in place technical measures to ensure Ukrainian civilians' physical safety. Russian military attacks on Ukrainian cities have severely impacted the civilian population. More than six million Ukrainians have had to flee abroad into Europe, and half a million beyond Europe, in addition to 3.7 million who are internally displaced (UNHCR 2024). In these circumstances, Big Tech corporations supported the efforts of government bodies and NGOs by playing the role of multipliers, that is, by promoting fundraising efforts amongst their employees and users. Additionally, they have provided their own funding to humanitarian aid operations, including those run by international governmental organizations, such as UN agencies, and NGOs, such as the Red Cross. For example, in March 2022, Microsoft committed USD 35 million towards humanitarian assistance projects in Ukraine, in support of NGOs such as the Polish Humanitarian Action and the International Red Cross (Endicott 2022). These funds contributed to operations supporting Ukrainians seeking safe spaces and shelters within Ukraine itself, as well as in neighbouring countries. Additionally, Google pledged USD 10 million to support humanitarian aid organizations working on short and long-term programmes for refugees in Poland (Alessandrini 2022). With regard to the initiatives taken on their platform specifically, Google.org and Google employees pledged USD 5 million to finance advertising contents by that disseminate information about resettlement, reputable humanitarian and intergovernmental organizations.

Along with in-kind donations, Big Tech companies have also exploited the technical features of their platforms to support humanitarian efforts and the localization of safe environments. On the technical level, social media platforms owned by Meta have introduced features that give more visibility to information concerning essential resources, such as housing and immigration assistance (Meta 2022). Facebook has cooperated with Red Cross societies and UN agencies to help Ukrainian users find information on the services of such organizations, including medical help and safe housing (Meta 2022). Similarly, the company has cooperated with the World Health Organization and the International Medical Corps to tailor the platform's Emotional Health Centre. Finally, Google launched an SOS alert on searches across Ukraine: when people searched for refuge and evacuation information, they received an alert pointing them to the United Nations' resources for refugees and asylum-seekers (Walker 2022). On Google Maps, the company has also added information on refugee and migrant centres in neighbouring countries (Walker 2022).

Thus, these initiatives can be divided into two main categories. The first category includes all initiatives that add resources, such as the provision of additional monetary resources to humanitarian aid organizations. Evidently, the provision of monetary support is not unique to Big Tech companies, but they enjoy particular outreach capabilities that can considerably amplify their contributions. As for highlighting information on social media or other services – such as Google Maps flagging information on services and aid to refugees – it is a type of contribution unique to such platforms. This second category includes actor-specific or actor-unique added value and contribution to civilian protection.

From a preventive perspective, the protection of civilians' privacy in countries affected by conflicts has been largely overlooked in both practitioners' and scholarly debates. Yet innovative technologies are increasingly used to determine the physical areas where military attacks would have the largest impact (Yaacoub et al. 2020). These include technologies used to retrieve and exploit data on the density of population or categories of population in a specific area. Social media companies own a variety of data about their users, including information on their location, nationality, gender and age, all often made visible on their platforms. Therefore, understanding how tech corporations can safeguard their users' privacy in conflict zones is relevant for civilian protection in the Ukrainian conflict, as well as in the context of contemporary warfare, where technology plays an important role in planning and executing attacks.

With regard to social media platforms, Facebook has implemented measures that allow its Ukrainian users to lock their profiles and hide their followers. This means that Ukrainian users' private accounts or their followers cannot readily be searched by Russian military and pro-Russia groups. In addition, Google has suspended the live traffic functionalities of Google Maps in Ukraine to limit the vulnerability of users to Russian attacks. It has also assisted the Ukrainian government and civilians in crisis prevention by creating an air raid alert to protect citizens against Russian air bombing. Only tech companies have the tools and powers to put in place such measures through their platforms, with significant outreach capabilities. Against the backdrop of recent initiatives by tech corporations to shape international norms on cyber hostilities, such as Microsoft's Digital Geneva Convention proposal (Jeutner 2019; Sutherland et al. 2015), the cases show that Big Tech companies also have the technical power to implement measures to protect civilians' physical safety by enhancing their digital privacy and by supporting international governmental and NGOs to support access to essential services.

4. CONCLUSION AND RECOMMENDATIONS

The use and impact of malicious cyber operations during the Russo-Ukrainian conflict have stimulated scholarly and policy debates over the significance of cyber operations in contemporary warfare. Evidently, the emergence of the cyber domain challenges existing resilience and deterrence frameworks. Opponents use the cyber domain mainly to seek competitive advantages and amplify the impact of military attacks that impose costs upon adversaries. The development and deployment of cyber capabilities have therefore become core aspects of geopolitical rivalries. Furthermore, cyber operations can harm national security in peacetime, as well as in moments of crisis and during conflicts. In this context, this paper has examined the role of Big Tech corporations in enhancing or complementing state cyber capabilities.

Through the analysis of the case presented, this paper has shown that Big Techs have enhanced Ukrainian cyber resilience by providing support for countering cyber attacks against critical infrastructure, cyber espionage and disinformation campaigns. Furthermore, Big Techs have contributed to the protection of civilians and supported humanitarian aid operations during the conflict. Six critical areas of involvement and contribution have been identified and analysed where Big Techs' involvement has enhanced, complemented and/or filled gaps in government cyber resilience capacities. In some cases, Big Tech actors provided a unique contribution because of their expertise, capacities and capabilities, as well as ownership and operationalization of digital assets, including digital infrastructure, data centres or service platforms. They have added value to the Ukrainian war effort through the design and/or implementation of preventive and mitigating strategies for the protection of critical and digital infrastructure, government institutions and civilians. Therefore, Big Techs have played, and will continue to play, an active and strategically significant role during the conflict by presenting themselves as and acting as strategically relevant and necessary partners of the Ukrainian government.

The fact that governments are increasingly combining traditional and cyber forms of aggressive operations, including cyber attacks, espionage and informational campaigns, invites scholars to reassess the strategic role and significance of tech companies in international security. The findings have important scholarly and policy implications. From a scholarly perspective, they add to the limited body of international-relations literature investigating the role of companies in international affairs and politics. While not questioning the centrality of states in international relations, this paper aligns with the view that industry players should not be considered exclusively as states' subordinates. While states remain dominant players in international security, the cyber capacities and capabilities of tech companies complement those of state actors. The paper does not enter into how Big Techs influence states' decision-making or what motivations and interests underlie their activities. It does, however, recognize that states may need their services and that the outputs of techs' involvement can provide state actors with strategic advantages. In this regard, the paper has important policy implications.

While this paper has mapped six areas of contribution and impact to state-level cyber resilience that arise from public-private cooperation, the involvement of the tech industry in these areas must follow thorough and informed risk assessments. These are needed to capitalize on the unique capacities of this industry segment, while minimizing risks to government independence, autonomy, positioning, international stature and values. Tailored risk assessments, minimum requirements and monitoring schemes should be designed and implemented to prevent the involvement of tech companies from harming national security and interests. The risk areas identified here should be addressed by risk assessment exercises, monitoring strategies and explored by scholars. We suggest that risks belong to one of two categories. They can arise from tech companies' own (intentional) conduct or from the manipulation of tech companies by foreign adversaries.

Risk assessments should investigate the motivations of tech companies, their business practices and partnerships to assess potential disloyalty and avoid conflicting interests or the leaking of sensitive or strategic government information. Along the same lines, the risk of having a cuckoo in the nest needs to be considered, so that companies do not cooperate with other states in a manner that is inconsistent with or harmful to states' strategic objectives and values. The activities of tech companies may also be a liability when these companies and/or their assets are manipulated, infiltrated or sabotaged by foreign adversaries aiming at stealing, manipulating data, destroying assets or disrupting government function. Risk assessments should be far-reaching, addressing the safety of Big Techs' products and solutions, which may be a vector of harm following malicious manipulations during their design, construction or distribution. Finally, risk assessments should also integrate internal consistency checks, with two objectives. The first is to ensure that different policy instruments, including those regulating the development and use of emerging and disruptive technologies, are coordinated, to avoid regulatory gaps. The second is to ensure that independent decision-making, democratic processes and fundamental principles are respected. Finally, while this paper has focused on five companies ('The Big Five'), future research could include other tech giants, namely Nvidia and Tesla; non-US tech companies; and smaller companies, in order to provide a broader overview of private companies' cooperation with states in cyber resilience.

To conclude, this case study examined, and other significant cyber attacks have shown, the implications of cyber attacks on national and international security. In the current international security landscape, tech companies play an increasingly significant role in the protection of national security, in peacetime and during crises and conflicts. States' cooperation with this industry segment and their increasing dependence on private services for national and international security can have beneficial as well as harmful effects on states' power and independence. The assessment of the risks related to public-private tech cooperation in national and international security should be integrated into risk assessments and should be the object of further research.

REFERENCES

Abels, Joscha. 2024. 'Private Infrastructure in Geopolitical Conflicts: The Case of Starlink and the War in Ukraine'. *European Journal of International Relations* 30 (4): 842–66.

Alessandrini, Sara. 2022. 'Google Will Use Office Space in Poland to Support Ukrainian Refugees'. CNBC, 7 March. https://www.cnbc.com/2022/03/07/google-will-use-office-space-in-poland-to-support-ukrainian-refugees.html.

Andrew, James and Kenneth Geers. 2015. '"Compelling Opponents to Our Will": The Role of Cyber Warfare in Ukraine'. In *Cyber War in Perspective: Russian Aggression Against Ukraine*, 39–48. NATO Cooperative Cyber Defence Centre of Excellence.

Aviv, Itzhak and Uri Ferri. 2023. 'Russian–Ukraine Armed Conflict: Lessons Learned on the Digital Ecosystem'. *International Journal of Critical Infrastructure Protection* 43: 100637.

Babic, Milan, Jan Fichtner and Eelke M. Heemskerk. 2017. 'States Versus Corporations: Rethinking the Power of Business in International Politics'. *The International Spectator* 52 (4): 20–43.

Beecroft, Nick. 2022. 'Evaluating the International Support to Ukrainian Cyber Defense'. Carnegie Endowment, 3 November. https://carnegieendowment.org/research/2022/11/evaluating-the-international-support-to-ukrainian-cyber-defense?lang=en.

Cerulus, Laurens. 2022. 'Ukraine's Digital Minister Pleads with Big Tech to Pressure Moscow'. *Politico*, 26 February. https://www.politico.eu/article/ukraine-russia-google-youtube-apple-and- netflix-facebook-digital-minister-mykhailo-fedorov-big-tech/.

Computer Emergency Response Team of Ukraine (CERT-UA). 2024. 'Плани UAC-0133 (Sandworm) щодо кібердиверсії на майже 20 об'єктах критичної інфраструктури України'. Published 19 April. https://cert.gov.ua/article/6278706.

Computer Emergency Response Team for the European Union Institutions, Bodies and Agencies (CERT-EU). 2024. 'Russia's War on Ukraine: One Year of Cyber Operations'. https://cert.europa.eu/static/threat-intelligence/TLP-CLEAR-CERT-EU-1YUA-CyberOps.pdf.

Culliford, Elizabeth. 2022. 'Facebook Owner Meta Will Block Access to Russia's RT, Sputnik in EU'. Reuters, 28 February. https://www.reuters.com/business/media-telecom/facebook-owner-meta- will-block-access-russias-rt-sputnik-eu-2022-02-28/.

Culliford, Elizabeth and Sheila Dang. 2022. 'Facebook, Instagram Globally Demoting Posts from Russian State Media – Meta'. Reuters, 1 March. https://www.reuters.com/technology/facebook- owner-meta-says-it-is-globally-demoting-posts-russian-state-media-2022-03-01/.

Dave, Paresh. 2022. ‘Exclusive: Google Blocks RT, Sputnik from Play App Store in Europe’. Reuters, 2 March. https://www.reuters.com/technology/exclusive-google-blocks-rt-sputnik-play-app-store-europe-2022-03-02/. Delerue, François. 2020. *Cyber Operations and International Law* 146. Cambridge University Press.

Endicott, Sean. 2022. ‘Microsoft Has Committed Over $35 Million to Help Ukraine’. Windows Central, 23 March. https://www.windowscentral.com/microsoft-has-committed-over-35-million-help-ukraine.

Erlich, Aaron and Calvin Garner. 2023. ‘Is Pro-Kremlin Disinformation Effective? Evidence from Ukraine’. *The International Journal of Press/Politics* 28 (1): 5–28.

EuRepoC. 2024. ‘Advanced Persistent Threat Profile: ATP28 – Exploiting Democratic Vulnerabilities in Cyberspace’. https://eurepoc.eu/wp-content/uploads/2023/07/APT28-EN.pdf.

Geers, Kenneth, ed. 2015. *Cyber War in Perspective: Russian Aggression Against Ukraine*. NATO Cooperative Cyber Defence Centre of Excellence.

Geppert, Mike and Christoph Dörrenbächer. 2014. ‘Politics and Power Within Multinational Corporations: Mainstream Studies, Emerging Critical Approaches and Suggestions for Future Research’. *International Journal of Management Reviews* 16 (2): 226–44.

Golovchenko, Yevgeniy, Mareike Hartmann and Rebecca Adler-Nissen. 2018. ‘State, Media and Civil Society in the Information Warfare Over Ukraine: Citizen Curators of Digital Disinformation’. *International Affairs* 94 (5): 975–94.

Google. 2024. ‘Protecting Free Expression from digital attacks’. https://projectshield.withgoogle.com/landing.

Huntley, Shane. 2022. ‘An Update on the Threat Landscape’. Google. Published 7 March. https://blog.google/threat-analysis-group/update-threat-landscape-ukraine/.

Jacobsen, Jeppe T. 2021. ‘Cyber Offense in NATO: Challenges and Opportunities’. *International Affairs* 97 (3): 703–20.

Kello, Lucas. 2013. ‘The Meaning of the Cyber Revolution: Perils to Theory and Statecraft’. *International Security* 38 (2): 7–40.

Kello, Lucas. 2017. *The Virtual Weapon and International Order*. Yale University Press.

Lange-Ionatamishvili, Elina, Sanda Svetoka and Kenneth Geers. 2015. *Strategic Communications and Social Media in the Russia-Ukraine Conflict*. NATO Strategic Communications Centre of Excellence.

Lewis, James A. 2022. *Cyber War and Ukraine*. Center for Strategic and International Studies (CSIS).

Maschmeyer, Lennart. 2021. ‘The Subversive Trilemma: Why Cyber Operations Fall Short of Expectations’. *International Security* 46 (2): 51–90.

Meta. 2022. ‘Meta’s Ongoing Efforts Regarding Russia’s Invasion of Ukraine’. Published 26 February. https://about.fb.com/news/2022/02/metas-ongoing-efforts-regarding-russias-invasion-of-ukraine/.

Microsoft. 2022a. ‘Digital Technology and the War in Ukraine’. Published 12 February. https://blogs.microsoft.com/on-the-issues/2022/02/28/ukraine-russia-digital-war- cyberattacks/?preview_id=65075.

Microsoft. 2022b. ‘Defending Ukraine: Early Lessons from the Cyber War’. Published 22 June. https://blogs.microsoft.com/on-the-issues/2022/06/22/defending-ukraine-early-lessons-from-the-cyber-war/.

Microsoft. 2022c. ‘Ukrenergo: We Couldn’t Survive Without the Cloud’. Published 12 December. https://news.microsoft.com/en-cee/2022/12/12/ukrenergo-we-couldnt-survive-without-the-cloud/.

Microsoft. 2023. ‘How Technology Helped Ukraine Resist During Wartime’. Published 20 January. https://blogs.microsoft.com/on-the-issues/2022/06/22/defending-ukraine-early-lessons-from-the-cyber-war/.

Ministry of Digital Transformation of Ukraine. 2024. ‘Strengthening Cyber Defence: Google Provides Ukrainian Civil Servants with 5,000 Security Keys to Protect Their Accounts’. Published 16 January. https://www.kmu.gov.ua/en/news/posyliuiemo-kiberzakhyst-google-nadaie-ukrainskym- derzhsluzhbovtsiam-5-tysiach-kliuchiv-bezpeky-dlia-zakhystu-oblikovykh-zapysiv.

Mozur, Paul and Adam Satariano. 2024. ‘Russia, in New Push, Increasingly Disrupts Ukraine’s Starlink Service’. *New York Times*, 24 May. https://www.nytimes.com/2024/05/24/technology/ukraine- russia-starlink.html.

Neuman, Noam. 2021. ‘Neutrality and Cyberspace: Bridging the Gap Between Theory and Reality’. *International Law Studies* 97 (1): 33.

Qamar, Sara, Zahid Anwar and Mehreen Afzal. 2023. ‘A Systematic Threat Analysis and Defence Strategies for the Metaverse and Extended Reality Systems’. *Computers & Security* 128: 103127.

Rid, Thomas. 2012. ‘Cyber War Will Not Take Place’. *Journal of Strategic Studies* 35 (1): 5–32.

Robinson, Michael, Kevin Jones and Helge Janicke. 2015. ‘Cyber Warfare: Issues and Challenges’. *Computers & Security* 49: 70–94.

Schulze, Matthias and Mika Kerttunen. 2023. *Cyber Operations in Russia’s War Against Ukraine: Uses, Limitations, and Lessons Learned So Far*. SWP Comment No 23/2023.

Schwartz, Matthew. 2017. ‘Microsoft Battles Fancy Bear Hackers – With Lawyers’. BankInfoSecurity. Published 31 July. https://www.bankinfosecurity.com/microsoft-battles-fancy-bear-hackers-lawyers-a-10156.

Štrucl, Damjan. 2022. ‘Russian Aggression on Ukraine: Cyber Operations and the Influence of Cyberspace on Modern Warfare’. *Contemporary Military Challenges / Sodobni Vojaški Izzivi* 24 (2): 103–23.

Sutherland, Iain, Konstantinos Xynos, Andrew Jones and Andrew Blyth. 2015. ‘The Geneva Conventions and Cyber-Warfare: A Technical Approach’. *The RUSI Journal* 160 (4): 30–39.

United Nations High Commissioner for Refugees (UNHCR). 2024. ‘Ukraine Situation Flash Update #73’. Published 25 September. https://data.unhcr.org/en/documents/details/111432.

USAID. 2022. ‘Cybersecurity’. National Security Archive. Published May. https://nsarchive.gwu.edu/sites/default/files/documents/rkbxys-4mwer/049-USAID-Cybersecurity-Fact-Sheet-May-2022.pdf.

van Benthem, Tsvetelina J. 2023. ‘Privatised Frontlines: Private-Sector Contributions in Armed Conflict’. In *2023 15th International Conference on Cyber Conflict: Meeting Reality (CyCon)*, 55–69. IEEE.

Venafi. 2022. ‘The (Nation) State of Cyber: 64% of Businesses Suspect They’ve Been Targeted or Impacted by Nation-State Attacks’. https://venafi.com/blog/nation-state-cyber-64-businesses-suspect-theyve-been-targeted-or-impacted-nation-state-attacks/.

Walker, Kent. 2022. ‘Helping Ukraine’. Google. Published 4 March. https://blog.google/inside-google/company-announcements/helping-ukraine/.

Willett, Marcus. 2022. ‘The Cyber Dimension of the Russia–Ukraine War’. In Survival: *October–November 2022*, 1st ed., edited by The International Institute for Strategic Studies (IISS). Routledge. https://doi.org/10.4324/9781003422211.

Yaacoub, Jean-Paul, Hassan Noura, Ola Salman and Ali Chehab. 2020. ‘Security Analysis of Drones Systems: Attacks, Limitations, and Recommendations’. *Internet of Things* 11: 100218.

CyCon 2025: The Next Step
17th International Conference on Cyber Conflict
C. Kwan, N. Gratzer, K. Podiņš, M. Tolppa (Eds.)
2025 © NATO CCDCOE Publications, Tallinn

Permission to make digital or hard copies of this publication for internal use within NATO and for personal or educational use when for non-profit or non-commercial purposes is granted provided that copies bear this notice and a full citation on the first page. Any other reproduction or transmission requires prior written permission by NATO CCDCOE.

The Need for Speed: Leveraging Civilian Contributions in a Rapidly Evolving Cyber Conflict

Gabrielle Joni Verreault
PhD Candidate in Bioethics
School of Public Health, Université de Montréal
Montreal, Canada
gabrielle.veilleux-verreault@umontreal.ca

Abstract: Cyberspace is an increasingly contested environment that includes new forms of inter- and intra-state conflict, such as industrial espionage, infrastructure hacking, disinformation, and election manipulation. Involvement in cyber operations is not limited to state or quasi-state actors but also includes civilians, thus challenging traditional ethical and legal frameworks such as the laws of war and armed conflict. Combining principles from bioethics and military ethics with empirical methodologies, a preliminary open-source ethical framework is presented to help guide civilian volunteer engagement in conflict zones. Drawing on empirical data from the ongoing conflict in Ukraine, the framework addresses the unique ethical challenges posed by civilian participation in cyber and hybrid warfare, spaces where identities, roles, and responsibilities can become blurred. The framework emphasizes adaptability, leveraging the concept of a "learning organization" (i.e., dynamic bottom-up and top-down co-development) to ensure that guidelines to orient civilians remain relevant amidst rapidly changing technological and geopolitical contexts. An open-source innovation approach is mobilized to foster a community-driven and continuously evolving structure that can be easily shared and adapted to diverse conflict environments, thereby enhancing resilience and responsiveness to the ethical complexities of decentralized civilian involvement. The aim is to provide civilians with structured yet flexible guidelines to safely navigate their various roles and responsibilities. The effectiveness of the framework will be analyzed using real-world case studies (e.g., drones, OSINT, hacktivism) from Ukraine, illustrating how civilian contributions could be ethically managed without compromising operational security or humanitarian protections. Policy recommendations are proposed to integrate and formalize the framework in an established organization, enabling a more robust, ethical approach to civilian cyber operations. A path forward is offered for policy-

makers, technologists (and ethicists!) to navigate the "next steps" in cyber conflict with greater clarity and ethical foresight.

Keywords: *ethics, framework, civil engagement, cyber conflict, Ukraine*

1. INTRODUCTION

The evolving nature of modern warfare, recent advances in both military and civilian technologies (e.g., commercial drones), and the expansion of cyberspace into many aspects of our lives have blurred the lines between combatants and non-combatants. International humanitarian law (IHL) aims to protect civilians during armed conflicts by making it unlawful to target them for hostile attacks. Those civilians directly participating in hostilities lose this protection, are considered "combatants," and can thus be legally targeted. If it is impossible to differentiate between those who are undertaking military activities and those who are not, IHL prescribes that they be presumed to have civilian status and be protected. The challenge, however, is in differentiating activities that are "hostile" or "combat" from those that are "simply" support. As cyberspace continues to grow in importance as the fifth dimension of warfare, it is transforming the nature of conflict. In a marked shift from state-sponsored cyber operations, civilians are increasingly active in hacktivism and using online platforms to organize and fundraise for causes they support, making them integral players in the domain. When and in what contexts do civilians become combatants when the terrain is cyberspace? And is the IHL sufficient legal and ethical protection for civilians?

To be clear, civilians have always played various supporting roles in conflicts. During World War II, in Canada, their involvement was called the "home front,"[1] with part of civil society mobilized to support the country's infrastructures or military campaigns. And in the United Kingdom, the Air Raid Precaution program mobilized civilian volunteers in response to German air raids.[2] Recently, some countries have adopted "total defense" policies to prepare society for crisis, the aim being to enhance national security and societal resilience by coordinating efforts across public, private, and civic sectors toward a common goal: survival.[3] IHL played a crucial protective role in these examples and is carefully crafted with accountability mechanisms for

[1] Jack L. Granatstein, "Wartime Home Front," *The Canadian Encyclopedia*, February 7, 2006, https://www.thecanadianencyclopedia.ca/en/article/wartime-home-front.

[2] Royal Air Force Museum, "Air Raid Shelters," Royal Air Force Museum, accessed January 6, 2025, https://www.rafmuseum.org.uk/research/online-exhibitions/history-of-the-battle-of-britain/air-raid-shelter-protection/.

[3] James Kenneth Wither, "Back to the Future? Nordic Total Defence Concepts," *Defence Studies* 20, no. 1 (December 2, 2019), https://doi.org/10.1080/14702436.2020.1718498.

belligerents who violate these protections. The civilians who remain outside combat zones are the easiest to protect, both physically and legally; they attend to their own well-being and that of their families. It becomes more complicated when some stay behind as volunteers, introducing the "fog of war"—the uncertainty faced by those in conflict —into civil society.

Today, volunteering is not limited to conflict zones or even national borders. Technology and increased connectivity allow civilians to engage in new ways as the internet opens up new possibilities.[4] This changing environment was first mentioned in the 2008 International Review of the Red Cross[5] and confirmed[6] in 2023 with Russia's full-scale invasion of Ukraine, where the traditional roles of non-combatants are evolving to include active participation in different capacities.[7] Aided by technologies that, among other things, allow civilians to self-organize, volunteers in Ukraine are playing roles in covert cyber operations, do-it-yourself tech support, and frontline logistics, all of which are also shared widely on social media. Cyberspace provides virtual public squares where civilians can engage individually and collectively, leveraging online spaces to advance social and moral objectives alongside coordination of their technical contributions. This holistic engagement in Ukraine enabled rapid mobilization and the scaling of efforts, encouraged grassroots humanitarian initiatives, and nurtured creative innovations to address immediate and long-term challenges.

While creating new opportunities and promoting innovation, this decentralization also heightens the risks for volunteers. Notably, they may inadvertently put themselves or others at risk of harm or undermine broader military operations and diplomatic efforts through uncoordinated or ethically questionable actions. It is thus imperative to establish a comprehensive ethical framework that can guide civilians' involvement in technical and civic dimensions, help them analyze the risks of particular choices or actions, and ensure that their contributions are ethically justified, if not legally protected, as the current legal frameworks are insufficient. This article proposes such a framework. Guided by the principle of harm reduction—recognizing that people will become involved regardless of attempts to dissuade them—states and the broader international community have the responsibility to help guide civilians to act as responsibly and securely as possible. In line with this approach, the proposed framework strives to balance the need for effective military support with protecting

[4] Peter Evans-Greenwood, "The Real Landscape of Technology-Enabled Opportunity," *Deloitte Insights*, February 28, 2022, https://www2.deloitte.com/us/en/insights/topics/innovation/technology-opportunity-landscape.html.

[5] Andreas Wenger and Simon J. A. Mason, "The Civilianization of Armed Conflict: Trends and Implications," *International Review of the Red Cross* 90, no. 872 (December 2008): 835–52, https://doi.org/10.1017/S1816383109000277.

[6] Kubo Mačák, "Will the Centre Hold? Countering the Erosion of the Principle of Distinction on the Digital Battlefield," *International Review of the Red Cross* 105, no. 923 (August 2023): 965–91, https://doi.org/10.1017/S1816383123000152.

[7] Marta Kepe and Alyssa Demus, "Resisting Russia: Insights into Ukraine's Civilian-Based Actions During the First Four Months of the War in 2022" (RAND Corporation, August 15, 2023), https://www.rand.org/pubs/research_reports/RRA2034-1.html.

civilian lives. Further, it recognizes the strategic opportunity to be leveraged by this new layer of civilian involvement. Given the limitations of current IHL in addressing the complexity and multidimensional nature of modern conflicts, it will be critical to establish an easily accessible platform or structure in which volunteers can operate safely, grounded in universally applicable principles that can guide civilian participation while not missing out on the values they bring to military operations.

2. THE UKRAINIAN EXAMPLE

In Ukraine, civilians have been the cornerstone of the ongoing war effort. Reflecting the country's dire need for resources and support, Ukrainian civil society mobilized and has shown many examples of what civilian involvement will look like in future conflicts.

The face of this wartime volunteering is threefold. First, traditional engagement has focused on addressing essential needs, with volunteers providing emergency medical care and evacuation to the injured on the front line.[8] Other civilians contribute by cooking (and then dehydrating) homemade borsch for the troops,[9] offering both sustenance and the psychological comfort of a traditional meal. Some volunteers take turns weaving camouflage nets for the army,[10] a resource that is widely needed nationwide to hide all sorts of equipment from observation by now-ubiquitous surveillance drones.

The second aspect highlights how some classic roles have been transformed by technology. Civilian "fixers," who have long been essential to foreign journalists by helping them to navigate local cultural differences and facilitate access to newsworthy contacts involved in the conflict,[11] have embraced the internet. Leveraging aggregator websites and social media platforms, Ukrainian fixers advertise their services online to connect with international journalists. Technology has also changed how fundraising activities are conducted, which has shifted from global NGOs to local groups now relying primarily on social media. The severe funding needs create a *Hunger Games*-type dynamic, where going viral or gaining popularity attracts sponsors and drives

[8] Mark Mansfield, "One of Ukraine's Most Popular Bands Visit Cardiff to Raise Money for Medical Aid," Nation. Cymru, June 1, 2024, https://nation.cymru/news/one-of-ukraines-most-popular-bands-visit-cardiff-to-raise-money-for-medical-aid/.

[9] "A British Volunteer in Ukraine Shares His Experience at Front Line Kitchen in Lviv," Volunteering Ukraine, May 23, 2023, https://www.volunteeringukraine.com/en/post/foreign-volunteer-in-ukraine.

[10] Isabelle Khurshudyan, "Sewing Camouflage in Kyiv: Women Volunteers Craft Cover for Ukraine's Military," *Washington Post*, February 8, 2022, https://www.washingtonpost.com/world/2022/02/08/ukraine-military-women-camouflage/.

[11] "I Didn't Choose to Be a Fixer, the Job Chose Me," DW Akademie, March 22, 2023, https://akademie.dw.com/en/fixers-in-ukraine-i-didnt-choose-to-be-a-fixer-the-job-chose-me/a-64842525; "Fixers: The Unsung Heroes of the War in Ukraine," Free Press Unlimited, September 1, 2022, https://www.freepressunlimited.org/en/current/fixers-unsung-heroes-war-ukraine.

donations for equipment needs for community groups and even individual military units.

The third aspect demonstrates how a life intertwined with technology can create new volunteering opportunities that could not have existed otherwise. Some volunteers have taken on roles as cyber combatants, receiving directives from the Ukrainian IT Army, which orchestrates cyberattacks against Russian organizations. Consisting of thousands of volunteers worldwide, hackers use Telegram channels to coordinate operations ranging from data theft to disruption of Russian communication networks.[12] Additionally, trolls have emerged as critical players in information warfare on social media.[13] Beyond agitating in the comment sections of Russian-controlled accounts, trolls can counter disinformation, shape pro-Ukraine narratives, challenge pro-Russian propaganda, and play an important role in fundraising for the Ukrainian military.[14] Lastly, the success of modern off-the-shelf commercial drones on the front line hinges on the civilians who possess the essential knowledge and skills to make this new war industry work. Civilians fund and order massive quantities of drones, physically modify them, and hack their firmware to make them frontline-ready. They also teach the military how to operate these drones, making civilian expertise crucial for military effectiveness. This civilian involvement has kickstarted a thriving, government-supported, low-cost, innovative, decentralized, and highly effective drone industry in Ukraine, making the country a global leader in the domain.[15] Even the most classic volunteer roles—the cooking mentioned above and "onsite" volunteering—have transformed with global connectivity and the spread of social media, and they all influence each other. This revolution has brought two notable changes: self-organization and remote involvement.

Ukrainians have suffered multiple war crimes and treaty violations, as well as energy shortages due to the destruction of infrastructure.[16] Those in unoccupied areas regularly hear of the Russian Armed Forces (RuAF) killing, confining, torturing, and raping their compatriots in the occupied areas, as well as about the systematic

[12] Vasileios Karagiannopoulos, "Ukraine's IT Army Is a World First: Here's Why It Is an Important Part of the War," *The Conversation*, October 25, 2023, http://theconversation.com/ukraines-it-army-is-a-world-first-heres-why-it-is-an-important-part-of-the-war-212745.

[13] Kathleen McInnis, Seth G. Jones, and Emily Harding, "NAFO and Winning the Information War: Lessons Learned from Ukraine," Center for Strategic and International Studies, October 5, 2022, https://www.csis.org/analysis/nafo-and-winning-information-war-lessons-learned-ukraine.

[14] Michael Drummond, "Ukraine's Internet Army of 'Fellas' Buy Sea Drone to Hunt Russian Ships—and Name It After Celebrity Raccoon," Sky News, January 24, 2023, https://news.sky.com/story/ukraines-internet-army-of-fellas-buy-sea-drone-to-hunt-russian-ships-and-name-it-after-celebrity-raccoon-12762265; Jason Jay Smart, "Founder of NAFO Reveals Identity, Discusses Raison D'être," *Kyiv Post*, November 14, 20222, https://www.kyivpost.com/post/204.

[15] Peter Dickinson, "Ukraine's Innovative Drone Industry Helps Counter Putin's War Machine," *Atlantic Council* (blog), June 26, 2024, https://www.atlanticcouncil.org/blogs/ukrainealert/ukraines-innovative-drone-industry-helps-counter-putins-war-machine/.

[16] "Bombardment of Ukraine's Power Generation and Transmission Infrastructure, 1 October 2022 to 30 April 2023: A Remote Assessment," Conflict Observatory, accessed June 16, 2024, https://hub.conflictobservatory.org/portal/apps/sites/?#/home/pages/power-1.

deportation of Ukrainian children to Russia.[17] These asymmetries have generated a rightful collective sense of injustice that has compelled civil society to rise to a new level of commitment toward volunteering, where the primary objective is to both stay alive and maintain the country's sovereignty in the face of this adversity.[18] They cannot correct these asymmetries at the source because they have no direct means of resisting many of the RuAF's actions. Nor can their country receive NATO membership and its resources while at war. Thus, Ukrainian (and other) civilians have responded by innovating to support their military's efforts on various fronts, such as fully leveraging the opportunities of cyberspace.

From within cyber environments, fighting disinformation, taking part in covert or state-backed cyber operations in urban areas, and modifying drones and teaching military personnel how to operate them in various environments, Ukrainians have shown the world what civilian engagement looks like in the 21st century. However, they are also paying, sometimes with their lives, for challenging their protected status under IHL: to stay competitive in an environment where their opponent is violating established rules, civilians may have to bend a few themselves. It may be time to create a new model that allows for autonomy and flexibility, ensuring that civilians who wish to participate actively in conflicts have a fighting chance but without losing their legal protection, or at the very least, understand well what they are risking through their engagement.

3. RISKS AND OPPORTUNITIES

Civilians are fully involved in this fifth domain and will likely remain integral to humanitarian and military efforts in future conflicts. It thus becomes critical to examine the implications of redefining their role while protecting them from consequences they may not have fully considered. However, civilians also represent an opportunity for state actors to leverage. Their motivation is best summarized by a sentiment often heard in Ukraine: "I want to help, but I would be useless with a gun." By letting them participate in support roles, without direct involvement in combat, civilians contribute significantly to military efforts, but in so doing, they inadvertently put a target on their backs. The existing rigid binary of combatant/non-combatant is inadequate. It may be time to create a new, less restrictive categorization for civilians, allowing them autonomy in their choice to support a conflict and independence in how to best engage. Creating such a space requires addressing inherent risks while capitalizing

[17] OHCHR, "Report of the Independent International Commission of Inquiry on Ukraine" (Office of the United Nations High Commissioner for Human Rights, March 18, 2024), https://www.ohchr.org/sites/default/files/documents/hrbodies/hrcouncil/coiukraine/a-hrc-55-66-aev.pdf.

[18] Nataliia Stepaniuk, "Wartime Civilian Mobilization: Demographic Profile, Motivations, and Pathways to Volunteer Engagement Amidst the Donbas War in Ukraine," *Nationalities Papers*, 2022, https://doi.org/10.1017/nps.2021.82.

on the opportunities their newfound freedom of movement can bring to military strategies.

A. Danger Zones

1) Distinction Principle

A core tenet of IHL is the principle of distinction, which mandates that a differentiation be made between combatants and civilians to protect non-combatants. Cyber operations make it hard to apply the distinction principle, as it is virtually impossible to differentiate between combatants, non-combatants, civilians, and military personnel.[19] Civilians with technical expertise can engage in cyber operations that directly support military objectives without bearing arms or being part of a formal military structure. This dual role complicates the application of IHL, as traditional definitions of combatants are insufficient to encompass these new forms of participation.

2) Attribution Principle

The cyber domain also exacerbates the attribution problem under *jus ad bellum*, the principle governing the conditions under which states may use force against one another. The prohibition against using force (except in self-defense) becomes problematic in the case of cyberattacks, because these can be carried out by independent or state-sponsored groups that are easily disavowed by the states that commission these actions. This then makes attribution and state retaliation complicated to justify.[20] Cyberspace's inherent anonymity and decentralized architecture enable individuals or loosely affiliated groups to conduct operations without establishing clear lines of accountability or necessitating authorization through a formal chain of command.

3) Retaliation and Proportionality

Cyberattacks also create retaliation problems, as it is not clear what constitutes an adequate and proportional response.[21] Cyberattacks can range from low-impact disruptions to high-stakes breaches of critical infrastructure, each carrying different levels of potential civilian harm. Responding proportionally to these threats is challenging because their effects can be widespread and difficult to predict, increasing the risk of unintended collateral damage. The nature of cyber retaliation is also ambiguous. While traditional kinetic responses are not impossible, they may be disproportionate or ineffective against cyber threats (especially when they are decentralized and conducted by civilians), necessitating the development of new forms of proportional responses that align with the unique dynamics of cyberspace.

[19] Robin Geiß and Henning Lahmann, "Cyber Warfare: Applying the Principle of Distinction in an Interconnected Space," *Israel Law Review* 45, no. 3 (November 2012): 381–99, https://doi.org/10.1017/S0021223712000179.

[20] Hensey Fenton, "Proportionality and Its Applicability in the Realm of Cyber Attacks," *Duke Journal of Comparative & International Law* 29, no. 2 (February 11, 2019): 335–59.

[21] Fenton, "Proportionality and Its Applicability in the Realm of Cyber Attacks."

B. The Gray Zone of Decentralization

The tensions above stem primarily from cyberspace's inherent characteristics: its borderless nature, anonymity, and the rapid pace at which operations can be initiated. However, its decentralized nature is what primarily brings both strength and risk to civilian volunteering. Most volunteers do not operate in active war zones but rather in urban centers untouched by constant military activity. In Ukraine's case, it is reasonable to ask whether volunteers' activities could draw Russian fire, given that their operations significantly reinforce Ukraine's military effectiveness. The RuAF is known not to limit strikes to the front lines; it frequently targets cities and non-military infrastructure hundreds of kilometers away. Many towns far from direct combat regularly experience attacks from drones, hypersonic missiles, and cruise missiles launched from as far as the Caspian Sea.[22] In these same urban areas, volunteers are engaged in defense support activities. The Russian Federation, already carrying out attacks under the pretext of targeting defense industries, could use the presence of volunteers as justification for further strikes. Setting aside the illegitimacy of Russian aggression, the notion of proportionality becomes paramount. An attack must be proportionate to the military advantage gained; this principle is essential to limit excessive harm to civilian society. While volunteers in defense industries contribute to military efforts, the proportionality of deploying high-impact weapons against decentralized grassroots targets is highly questionable, given that specific points of volunteer activity matter less than the volunteer effort does in the aggregate.

Turning to the cyber front, the challenges of legitimate targeting and proportionality have become even more complex, particularly since the rules of engagement are still evolving.[23] Hacktivists, often operating from urban centers away from the front lines (or in other countries), may be deemed combatants, raising similar questions of proportionality. A coordinated cyberattack by an official military body might theoretically justify a kinetic response—although this remains a subject of debate—but the same attack carried out by multiple, decentralized civilian groups hardly justifies a destructive response and may even make it impossible in practice. Striking the presumed location of a cyber operator in a populated city would likely incur severe civilian casualties and infrastructure damage. Moreover, the dispersed nature of cyber operations means their overall effectiveness stems from numerous small actions rather than a single large-scale attack from a single point. This fragmentation further complicates any justification for high-impact or indiscriminate responses.

Yet many examples in Ukraine show that even though these emerging civilian roles may increase risks, they can also contribute to new, unconventional tactics in

[22] "Russian Federation's War Having 'Appalling Impact' on Ukraine's Children, Under-Secretary-General Tells Security Council," Meetings Coverage and Press Releases, United Nations, January 10, 2024, https://press.un.org/en/2024/sc15559.doc.htm; Yelnur Alimova, "Russia Is Using the Caspian Sea to Launch Strikes Against Ukraine. So Why Are the Caspian Countries Silent?," Radio Free Europe/Radio Liberty, December 2, 2022, sec. Russia, https://www.rferl.org/a/caspian-sea-ukraine-war-russia-peace-friendship-convention/32158822.html.

[23] Geiß and Lahmann, "Cyber Warfare."

modern conflicts. Expert volunteers can be invaluable, and rejecting their help would be strategically unwise. At the same time, expanding civilian involvement carries responsibilities that should be understood. In moving toward a new framework to govern such civilian involvement, we must account for the complexities of cyber operations and urban defense by minimizing harm to civilians and empowering them to engage responsibly. Integrating civilian volunteers introduces challenges familiar to military cyber operations, but the growing number of participants amplifies these issues. We must therefore consider holistic strategies that acknowledge civilians' capabilities and ensure that their contributions fit safely and ethically within the broader defense landscape.

C. Zone of Opportunities

1) Self-Organization

The nature of modern warfare increasingly involves civilians in indirect military activities, allowing them to volunteer their skills where they feel they could be helpful. Aided by technology, one of the key aspects of modern volunteering is self-organization. In Ukraine, this new trend is highlighted on social media.[24] There is no need to rely on large structures and organizations—civilians use democratized social platforms to organize their cyber operations, provide do-it-yourself tech support, or take part in logistics.[25] Cyberspace serves as a catalyst for this self-organization, altering the dynamics of conflict participation. These ad hoc, loosely organized networks can mobilize swiftly and operate autonomously, offering flexibility, scalability, and a fast response to emerging threats or opportunities within the conflict.

2) Nurturing Innovation

Decentralization gives volunteers independence and fosters a culture of innovation, as diverse groups bring varied skills and perspectives, enabling the development of novel solutions. Within these environments, civilians can apply their expertise and skills and pursue their interests, providing support from their comfort zones and contributing to the war effort. Civilian-led spaces democratize participation and allow individuals from different backgrounds to contribute based on their capabilities. This inclusivity broadens the talent pool and strengthens collective efforts by integrating diverse approaches and methodologies.

3) Resilience and Redundancy

Decentralization enhances resilience against targeted attacks in the context of hacktivism. A decentralized network is less vulnerable to single points of failure,

[24] Kateryna Fedotenko, "Cyber Warfare as Part of Information Warfare of Russia against Ukraine since the Beginning of the 2022 Russian Invasion," *Věda a Perspektivy*, no. 8(27) (August 27, 2023), https://doi.org/10.52058/2695-1592-2023-8(27)-351-357; Stepaniuk, "Wartime Civilian Mobilization."

[25] Kateryna Denysenko, Dmitri Kyseliov, and I. Borko, "Some Aspect of Volunteers Activities Under Conditions of Martial Law," *Scientific Herald of Sivershchyna. Series: Law*, 2023, https://doi.org/10.32755/sjlaw.2023.02.018.

making it harder for adversaries to dismantle civilian support efforts through conventional cyber or kinetic means. Decentralized systems inherently incorporate redundancy, allowing multiple nodes to operate independently and ensuring continuity of operations. This distributed approach mitigates the risk of identification or the impact of targeted attacks and facilitates rapid recovery and adaptation, thereby maintaining operational integrity under adverse conditions.

4. THE FRAMEWORK

The framework for civilian volunteers is anchored in the principles of two domains of applied ethics: bioethics and military ethics. Bioethics, emphasizing human well-being and ethical decision-making in challenging (including life-and-death) situations,[26] offers a holistic approach to understanding the complex moral issues facing civilians in conflict zones. Military ethics provides a pragmatic perspective, focusing on professional/institutional responsibilities and the values that make civilian actions meaningful in the context of military operations.[27] This ensures that their actions are morally justified and strategically sound. Thus, the proposed framework offers ethically grounded and practically applicable guidelines for civilians, helping them maintain moral integrity while effectively contributing to defense efforts. This framework, which could include a code of conduct, will be operationalized through comprehensive guidelines that civilians can understand and apply.

To positively shape civilians' involvement in conflicts, they should be guided towards safe practices and helped to recognize dangers they might overlook or underestimate while protecting military personnel who could be negatively affected by their participation. This means that the framework's ethical principles should reflect and represent civilian values. However, learning from the military is also crucial for two reasons. First, it is vital to understand how the growing number of civilians[28] supporting military efforts changes the dynamics of conflicts.[29] While their involvement can be beneficial, it can also be detrimental,[30] as civilians who have good intentions but who

[26] Kenneth Iserson, "Ethics of Emergency Department Cancer Care," 2016, 43–56, https://doi.org/10.1007/978-3-319-26387-8_3; Antonio Sandu, "The Principles of Bioethics and Their Use in Ethical Decision-Making," *Logos, Universality, Mentality, Education, Novelty. Section Social Sciences* IX, no. 1 (2020): 139–54.

[27] Paul Robinson, "Ethics Training and Development in the Military," *The US Army War College Quarterly: Parameters*, 2007, https://doi.org/10.55540/0031-1723.2344; "Virtue Ethics and Military Ethics," *Journal of Military Ethics* 6, no. 4 (2007): 257–58, https://www.tandfonline.com/doi/abs/10.1080/15027570701840455.

[28] Stepaniuk, "Wartime Civilian Mobilization."

[29] Olha Baidarova, "Features of the Volunteer Organizations Management in the Direction of Assisting the Military in Conditions of War," in *Sociology—Social Work and Social Welfare: Regulation of Social Problems* (Lviv, May 18–19, 2023): Proceedings of the XIII International Scientific Conference, January 1, 2023, https://www.academia.edu/103091563/features_of_the_volunteer_organizations_management_in_the_direction_of_assisting_the_military_in_conditions_of_war.

[30] Denysenko, Kyseliov, and Borko, "Some Aspect of Volunteers Activities Under Conditions of Martial Law."

lack appropriate training and discipline could unintentionally jeopardize operations.[31] Second, the military's structure, professional discipline, and organization can inspire the future framework, guiding civilians to behave appropriately, effectively, and even "professionally."

A. Ethical Backbone: Values and Principles That Drive Purpose

The framework is anchored in fundamental values that guide civilian volunteers' actions, keeping them ethically grounded in high-pressure situations. While these personal convictions align efforts with moral imperatives, relying solely on values can lead to emotional responses that obscure the most appropriate course of action.[32] Empathy plays a potent role, fueling moral development through its affective (feeling), cognitive (understanding), and motivational (desire to act) dimensions.[33] Driven by empathy, volunteers may view their actions as moral imperatives rather than obligations.[34] Yet empathy also carries the risk of bias, which may produce slanted or irrational decisions.[35] To counterbalance such pitfalls, the framework integrates the concept of "well-being" as defined by the World Health Organization (WHO), emphasizing not just individual but collective quality of life.[36] This twofold approach—akin to "securing one's oxygen mask first"—allows volunteers to remain effective without burning out.

Building on these foundations, the principle of duty introduces military and virtue ethics, instilling responsibility, competence, camaraderie, respect, courage, resilience, and discipline.[37] Guided by these virtues, civilian volunteers uphold accountability and reinforce cohesion, ensuring their actions bolster rather than compromise shared objectives.[38] Meanwhile, safety demands care for one's own physical and mental well-being, thorough preparation, and self-control. Collaboration naturally emerges from duty, centered on communication and mutual respect, so volunteers, military personnel, and communities can work collectively toward success.[39] Finally, the principle of viability ensures that volunteer contributions strive for long-term impact, avoiding short-lived fixes or dependency. By focusing on strategic planning and

[31] Jo E. Condrill, *Civilians in Support of Military Field Operations* (Carlisle Barracks, PA: U.S. Army War College, 1993), https://doi.org/10.21236/ada265397.

[32] Mary Frances Luce, James R. Bettman, and John W. Payne, "Choice Processing in Emotionally Difficult Decisions," *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23, no. 2 (March 1997): 384–405, https://doi.org/10.1037//0278-7393.23.2.384.

[33] Hannah Read, "A Typology of Empathy and Its Many Moral Forms," *Philosophy Compass* 14, no. 10 (2019): e12623, https://doi.org/10.1111/phc3.12623; Ted van Baarda and Désiré E. M. Verweij, *Military Ethics: The Dutch Approach: A Practical Guide* (Martinus Nijhoff Publishers, 2006).

[34] Baarda and Verweij, *Military Ethics*.

[35] Read, "A Typology of Empathy and Its Many Moral Forms."

[36] Rüdiger Krech et al., *World Health Organization. Health Promotion Glossary of Terms 2021*, (World Health Organization, 2021), https://iris.who.int/bitstream/handle/10665/350161/9789240038349-eng.pdf?sequence=1.

[37] Peer de Vries, "Virtue Ethics in the Military: An Attempt at Completeness," *Journal of Military Ethics* 19 (2020): 170–85, https://doi.org/10.1080/15027570.2020.1814048.

[38] Michael J. Zimmerman, "Duty and Obligation," in *The International Encyclopedia of Ethics*, ed. Hugh LaFollette (Chichester, UK: Wiley-Blackwell, 2013), https://doi.org/10.1002/9781444367072.WBIEE158.

[39] De Vries, "Virtue Ethics in the Military."

clear objectives, viability promotes sustainable interventions that endure beyond the immediate demands of a conflict.

B. Under the Hood: Conception and Methods

The framework must be specialized enough to serve its purpose but not so overspecialized that it becomes static and unusable in other contexts. Whether its normative and prescriptive outcomes apply to different conflicts depends on the methodological aspects and contextual elements that influence its broader applicability. The framework is envisioned not as a perfect and static solution but as a dynamic, learning-oriented system that evolves through continuous application and feedback from a community of practice made up of its users. Keeping the framework agile, documenting what works and does not, identifying context-specific elements, and refining its guidelines will keep it relevant in rapidly changing technological and geopolitical landscapes. Inspired by the concept of "learning organizations," emphasizing the importance of continuous adaptability and learning from all stakeholders in a changing environment, this approach is relevant to modern, fast-paced, and rapidly shifting conflict dynamics. Editable in nature, drawing on Agile principles,[40] the framework will be a living document that evolves through iterative feedback from real-world experiences. An important caveat of this approach is that it requires an active community of practice, which must adopt, maintain, and adapt the framework over time.[41] In response, the proposal is to build a virtual community of practice (VCoP), reminiscent of hacker culture,[42] in which communities operate and learn together, exchanging exploits and techniques.[43] Encouraging these traits to create a culture of collaboration would ensure that their involvement remains a resilient, innovative, and autonomous asset to conflicts. A VCoP working closely with national security questions and contexts must be supported by an organization (state or intra-state) dedicated to their best interests.

C. Architecture

Imagine a software developer sitting at home, far from any front line, and deciding to volunteer after watching troubling updates on social media. They log onto a digital platform—an "always-on hackathon"—where contributors from everywhere can work solo or team up to bring their expertise to conflict-related challenges. Upon signing up, individuals are presented with a thoughtfully developed code of conduct.

[40] Outi Salo and Pekka Abrahamsson, "An Iterative Improvement Process for Agile Software Development," *Software Process: Improvement and Practice* 12 (2007): 81–100, https://doi.org/10.1002/SPIP.305.

[41] Anne Hess, Philipp Diebold, and Norbert Seyff, "Towards Requirements Communication and Documentation Guidelines for Agile Teams," *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)*, 2017, 415–18, https://doi.org/10.1109/REW.2017.64.

[42] Leonie Tanczer, Irina Brass, and Madeline Carr, "CSIRTs and Global Cybersecurity: How Technical Experts Support Science Diplomacy," *Global Policy* 9 (2018): 60–66, https://doi.org/10.1111/1758-5899.12625.

[43] Baidyanath Biswas et al., "A Text-Mining Based Cyber-Risk Assessment and Mitigation Framework for Critical Analysis of Online Hacker Forums," *Decision Support Systems* 152 (2022), https://doi.org/10.1016/J.DSS.2021.113651.

This document functions more as an ethical agreement than a strict set of rules. It highlights the importance of responsible behavior and outlines legal boundaries, detailing volunteers' rights and the rights they relinquish when participating in activities that could affect ongoing military or relief operations, ensuring that they are informed before consenting to participate.

Once they accept these guidelines and enter the platform, volunteers can explore a live "community roadmap" featuring open projects, each with specific tasks such as patching a security vulnerability, upgrading drone firmware, conducting an OSINT (open-source intelligence) investigation into a potentially corrupt official, or organizing food and equipment deliveries to front-line combat units. Each project tile or listing shows the skills needed and how the effort contributes to the broader mission. Some volunteers may recognize that they already have the right background for a given task, while others may discover new challenges that spark their interest but demand fresh knowledge.

Newcomers then move into training and education modules designed with branching paths that reflect the diverse nature of conflict-related needs. Some prefer hands-on engineering or coding tutorials, while others delve into methods for social media monitoring, intelligence gathering, or countering disinformation. A novice might start with digital hygiene and foundational OSINT lessons, whereas a seasoned one can go straight into complex tutorials on ethical disruption. Just as important as competence is volunteer well-being. Each training path includes resources on stress management, avoiding burnout, and staying emotionally grounded. This balanced approach keeps volunteers healthier, resilient, and ultimately better equipped to handle the dynamic environment of conflict-driven projects.

Participants later move into the heart of the platform: a vibrant VCoP that blends real-time interaction with version-controlled documentation. Discussion channels reminiscent of Discord or Slack let volunteers plan projects, troubleshoot issues, or brainstorm together. Ideas flow freely but could be refined through a system akin to GitHub pulls and merges, ensuring that every proposal—whether a drone firmware patch or a new step-by-step guide to online fundraising—goes through a transparent review by senior members and VCoP moderators. A wiki-like repository captures best practices, creating a living knowledge base that remains visible and up-to-date. The open-source approach of the platform keeps it agile, allowing it to change quickly if, for instance, a new threat emerges or if a particular volunteer project shows signs of success and has methods that others might wish to replicate.

As the VCoP grows, it relies on feedback and iteration mechanisms designed for accountability without stifling creativity. An oversight body comprising military experts, humanitarian representatives, and trusted volunteer moderators regularly audits proposals and projects. These reviews reinforce the idea that even self-organized, decentralized volunteer efforts can have proper oversight without hampering creative freedom. This could mean a security professional reminding an overenthusiastic recruit why specific hacking methods or targets might violate international law and have unintended consequences, or a humanitarian coordinator sharing logistics-planning tips that include how best to organize shipping and to verify that donated medical supplies reach clinics instead of vanishing into black markets.

Within this perpetual hackathon, implementation strategies evolve to keep the platform open and secure. Community-driven development thrives on the principle that any participant can propose not only projects but also enhancements in the platform and can fix glitches or improve guidelines (e.g., context-specific standard operating procedures). This crowdsourced approach steadily evolves the platform through direct input from volunteers encountering daily issues. Moreover, continuous learning becomes second nature in the community. If an update on a project or new technique emerges, the platform's maintainers promptly revise the knowledge base, ensuring that others can access and use the latest information.

At the policy level, governments and international organizations can formalize this hybrid system of civilian participation by establishing a registration and vetting system that builds trust by respecting volunteers' privacy. Clear guidelines and a transparent sign-up process help recruits understand the potential risks and their responsibilities. But equally important are robust public awareness and information campaigns to demystify what it means to be a volunteer. These should clarify that volunteerism is not a free pass to engage in reckless hacking or misguided philanthropic efforts; it is a structured and ethically grounded commitment to aid a specific cause.

The success of this environment also depends on international cooperation and standardization. Aligning it with established institutions would ensure cohesive standards. This would mean that, no matter where a volunteer logs in from, they would find consistent baseline rules, better support systems, and a clearer understanding of operational protocols. This collaboration would secure a measure of legitimacy, making each contribution part of a global strategy rather than an isolated act of goodwill.

Finally, the framework includes a support system for volunteer well-being tailored to the intense nature of their engagement. Peer support chats, virtual "break rooms," and access to mental health resources help volunteers manage stress, particularly when they grapple with the moral weight of their chosen projects. When the going gets tough, or a crisis demands an urgent pivot in priorities, volunteers can lean on a community that understands these pressures and responds in real time.

A unique fusion of team-based problem-solving, open-source collaboration, and personal initiative emerges—an ongoing hackathon that any ethically minded civilian can join. By combining version-controlled documents, shared channels for brainstorming, and step-by-step training modules, this platform respects personal autonomy while inviting collective synergy as a vehicle towards acting in solidarity with a cause. It channels decentralized energies into cohesive, beneficial actions, ensuring that globally dispersed volunteers can work alone or side by side on projects that genuinely support conflict or humanitarian objectives. Through continual iteration, the platform becomes an innovation incubator supported by a living set of guidelines and a practical expression of global solidarity that can adapt to each new challenge and conflict.

5. CONCLUSION

One way to envision this framework is as a layered funnel, with each stratum catching different types of civilian volunteers and channeling them toward safer, more ethical engagement. The widest layer is IHL, which protects civilians who avoid participating in direct hostilities. From there, the proposed framework becomes the second layer, offering guidelines and oversight for those who wish to be more active and who are willing to adhere to a structured code of conduct. Even then, not everyone will opt in; some volunteers will remain on the fringe due to mistrust of formal structures or an insistence on autonomy. The framework proposes a minimal "cyber-IFAK" (individual first aid kit) anchored in a harm-reduction approach for that third and narrowest layer: a distilled packet of essential ethical and operational do's and don'ts. This ensures that even "rogue" volunteers have baseline guidance to mitigate the risk of harming themselves and those around them.

While this funnel-like approach underscores the framework's attention to a proportional approach to risk management, it also reveals the framework's inherent limits. It cannot and does not seek to force compliance. Instead, it provides entry points for civilian participants, whether they embrace complete oversight or stay independent. A volunteer may begin as a newcomer, integrate into a team project, and gain specialized skills and deeper knowledge of operational security, collaboration methods, and ethical constraints over time. At that peak of expertise, they might look to transition out of active involvement. Here, the framework adds an "exit strategy," guiding volunteers through a gentle off-ramp that debriefs them, helps them reorient, and offers referrals to new avenues, be this a peacetime role in civilian tech, a support system for mental well-being, or simply the formal recognition of their service. As such, the framework reduces the risk of burnout, disillusionment, and unintended radicalization by viewing civil volunteerism as a lifecycle that starts with a volunteer's entry, peak performance, and eventual reintegration into civilian life.

It is not a universal or perfect solution to the diverse ethical and security challenges experienced by civilians who volunteer in conflicts. Misuse remains possible, particularly by aggressors or extremist groups who might co-opt these resources for their own purposes, including to attack civilian volunteers. At the same time, the platform's open-source nature can serve as a powerful check: a transparent community can detect and call out deviations from humanitarian principles, while an innovative culture and a widely distributed community can help protect members from attack. Furthermore, the conversations supported by this framework extend beyond crises and armed conflicts. The volunteer culture cultivated here—rapid coordination, grassroots mobilization, and a commitment to ethical practice—can pivot to peacetime endeavors like disaster relief, community-led innovation, or public health emergencies. In that sense, we shift from "How do we harness this energy during war?" to "How do we channel it to respond to a public health crisis?"

Ultimately, this framework must be seen as an evolving call to action rather than a rigid rulebook. As the saying goes, "If you build it, they will come." But for the "building" to be trusted, it must show that it is effective while also offering layered safeguards, oversight, and a lifecycle approach to volunteer engagement and ensuring that civilian involvement in conflicts is as ethical, constructive, and future-proof as possible, even if some volunteers will always remain outside the framework's bounds. The multi-layered, iterative design promises to be able to adapt to new crises, reinforcing individuals' autonomy and promoting the collective imperative to protect life, minimize harm, and empower volunteers, regardless of the nature of the battlefield.

CyCon 2025: The Next Step
17th International Conference on Cyber Conflict
C. Kwan, N. Gratzer, K. Podiņš, M. Tolppa (Eds.)
2025 © NATO CCDCOE Publications, Tallinn

Permission to make digital or hard copies of this publication for internal use within NATO and for personal or educational use when for non-profit or non-commercial purposes is granted provided that copies bear this notice and a full citation on the first page. Any other reproduction or transmission requires prior written permission by NATO CCDCOE.

The Cloud of War: How Russian Military Mobile Applications Exploit Western Tech in the War Against Ukraine

Volodymyr Styran
State Cyber Protection Center
State Service of Special Communications and Information Protection
Kyiv, Ukraine

Abstract: This study conducts an analysis of Russian mobile applications deployed in the ongoing war of aggression against Ukraine, focusing specifically on their dependency on Western cloud and IT infrastructure. The investigation involved acquiring APK files of various apps used in military and intelligence operations, revealing a staggering prevalence of Western technology in their architecture. Key findings indicate that these applications extensively utilize Western cloud services for data storage, media streaming, and access management, alongside the global DNS services, anti-DDoS, and cybersecurity solutions to secure communication channels and protect against cyberattacks.

The study also uncovered significant reliance on global cloud service providers, which play a critical role in supporting the backend infrastructure of these apps. Furthermore, virtual private server (VPS) providers were identified as integral components in maintaining server operations and data processing for these war-related tools.

This widespread adoption of Western technological resources in applications directly tied to the military efforts of a state engaged in an aggressive invasion raises profound ethical and strategic concerns. It highlights a paradox in which companies from democratic nations—often nations imposing sanctions against Russia—inadvertently support the technical backbone of Russia's military operations. The study calls for a reevaluation of the policies governing the use of these platforms in conflict zones, emphasizing the need for stricter regulations and greater accountability from tech giants.

Keywords: *Russian mobile apps, war in Ukraine, cloud service providers, military technology, cyber warfare, digital geopolitics, technological dependencies, ethical implications, Western IT infrastructure*

1. INTRODUCTION

The war in Ukraine has highlighted the transformative power of technology in modern conflict. As Russian aggression continues, international technology companies have played a critical role in bolstering Ukraine's defenses, reshaping the battlefield in unprecedented ways.[1] Notable examples include SpaceX's Starlink, which has ensured uninterrupted connectivity amidst widespread infrastructure damage,[2] and Microsoft, whose cybersecurity efforts have shielded Ukraine's critical systems from persistent cyberattacks.[3] Additionally, companies like Palantir Technologies have provided advanced data analytics platforms, enabling Ukrainian forces to process vast amounts of information for strategic and tactical decision-making.[4]

These contributions underscore how technology, once a tool of convenience, has become a lifeline in wartime scenarios. However, this technological support has not been one-sided. Paradoxically, the very platforms that aid Ukraine also serve the adversary, creating a complex web of dependency and ethical dilemmas. Western cloud infrastructure and cybersecurity solutions, integral to Ukraine's defense, are simultaneously exploited by the Russian military.

Discussions under international humanitarian law (IHL) have raised concerns about technology companies supporting Ukraine's defense being considered military targets, given their role in aiding military operations.[5] However, there has been little to no discourse on the implications of these same companies inadvertently aiding the adversary. This oversight ignores critical questions about whether such involuntary support makes these companies complicit in potential violations of IHL and whether it imposes a duty to mitigate such risks.

[1] Diya Li, "On the Digital Front Lines: How Tech Companies Are Supporting Ukraine," U.S. Chamber of Commerce, March 29, 2022, https://www.uschamber.com/technology/on-the-digital-front-lines-how-tech-companies-are-supporting-ukraine.

[2] Dearbail Jordan, "Ukraine War: Elon Musk's Starlink System Helps Ukrainian Army Strike Russian Targets," BBC News, September 8, 2023, https://www.bbc.com/news/world-europe-66752264.

[3] Brad Smith, "Defending Ukraine: Early Lessons from the Cyber War," Microsoft, June 22, 2022, https://blogs.microsoft.com/on-the-issues/2022/06/22/defending-ukraine-early-lessons-from-the-cyber-war/.

[4] Vera Bergengruen, "AI in Ukraine War: How Palantir's Technology Helps in the Fight Against Russia," Time, February 8, 2024, https://time.com/6691662/ai-ukraine-war-palantir/.

[5] Jonathan Horowitz, "When Might Digital Tech Companies Become Targetable in War?," Tech Policy Press, October 13, 2023, https://www.techpolicy.press/when-might-digital-tech-companies-become-targetable-in-war/.

The lack of clear guidance on the ethical and legal responsibilities of global technology providers in conflict zones reveals a critical gap that demands urgent attention. This study seeks to bring this issue to the forefront of discourse, urging tech firms to recognize and address the abuse of their technologies. By highlighting the dual-use nature of these technologies, the study underscores the urgent need for a robust response to prevent their further exploitation in the prosecution of an illegal war.

2. METHODOLOGY

The research methodology employed in this study consisted of the following key steps:

1. Application Download and Dynamic Analysis
The first step involved searching, cataloguing, and downloading a set of military mobile applications and reviewing their features in an isolated, controlled environment. This dynamic analysis sought to confirm the military purpose of each app by observing functionality and data. This approach ensured that only apps with a confirmed military purpose were included in subsequent analyses.

2. Static Analysis for Network Data Extraction
Once the military purpose was confirmed, static analysis was performed on the applications. This involved decompiling the APK files to examine their code, configurations, and embedded resources. The goal was to extract networking data, including domain names, IP addresses, API endpoints, and other identifiers pointing to external resources and services. These data points revealed the infrastructure supporting the operational functionality of the apps.

3. Network Data Analysis
In the next stage, the extracted networking data was analyzed to identify the ownership and geographical distribution of the supporting resources. This included determining the cloud service providers, hosting platforms, and applications supplying the apps with essential data and services.

This approach provided an understanding of the technological dependencies of the apps, focusing on their military relevance and operational context. While minor inaccuracies may arise from obsolete or unused resources embedded in the apps, these are negligible and do not affect the validity of conclusions drawn from the significant volume of findings.

3. SAMPLE SELECTION

To examine the exploitation of Western technology in Russian military operations, a collection of 243 mobile applications was analyzed. These apps were identified as being used by Russian military forces on the battlefield. In the initial phase, civil, dual-use, and repurposed Ukrainian applications were excluded to narrow the focus to strictly military tools. This exclusion resulted in a final sample of 62 applications classified as strictly military. See Table I for the list of selected military apps.

The selected sample of strictly military mobile applications was categorized based on their primary functions. The largest categories, “Artillery Management” and “AUV Management,” included, respectively, 25 and 14 apps designed to assist with reconnaissance, targeting, and fire control. “Ballistic Calculators” and “Explosives Calculators” accounted for two apps each, providing precise computational tools for projectile trajectories and explosives parameters. “Field Logistics & Support” included four apps dedicated to managing supplies and operational logistics. “Mapping & Navigation” featured six apps, emphasizing geospatial awareness and navigation in tactical scenarios. Additionally, the sample contained tools for “Medical & Training,” “Meteorological Tools,” and “Tactical Communication,” showcasing a diverse array of functions critical to battlefield operations. See Table II for the list of identified military app categories.

TABLE I: RUSSIAN MILITARY APPS UNDER ANALYSIS

Category	App Name	Description
Artillery Management	120-note	Artillery app for managing 120 mm systems.
Artillery Management	122-note	Artillery app for managing 122 mm systems.
Artillery Management	152-note	Artillery app for managing 152 mm systems.
Artillery Management	2A80-note	Artillery app for managing 2A80 systems.
Artillery Management	2B11-note	Artillery app for managing 2B11 mortar.
Artillery Management	2B16-note	Artillery app for managing 2B16 Nona-K.
Artillery Management	2C4-notepad	Artillery app for managing 2C4 system.
Artillery Management	2S7 note	Artillery app for managing 2S7 Pion system.
Artillery Management	30-82-100	Artillery app for managing 30-82-100 mortar.
Artillery Management	Art-note	Versatile artillery fire control tool.

Artillery Management	BM21-note	Artillery app for managing BM-21 Grad.
Artillery Management	PUO-10E	A tool for artillery fire management.
Artillery Management	HM16 note	A tactical app for managing mortar artillery tasks.
Artillery Management	Hyacinth	Artillery app for Hyacinth system.
Artillery Management	M-46 note	Artillery app for M-46 field gun operations.
Artillery Management	Msta-note	Artillery app for Msta howitzer operations.
Artillery Management	Nona-note	Artillery app for 2C9 Nona operations.
Artillery Management	Spotter	Defining positions, calculating trajectories, and adjusting magnetic declination for accurate targeting.
Artillery Management	ZeVs—PZK	Assists in maintaining artillery performance by monitoring and calculating barrel wear over time.
Artillery Management	ZeVs—Artillery Grid	Supports artillery operations by enabling precise targeting and adaptability to environmental conditions.
Artillery Management	ArtGruppa	Part of the Veterok-ArtGruppa tactical software suite for reconnaissance and artillery units.
Artillery Management	ArtSkill	Fire adjustment training and test app.
Artillery Management	Armor	Tactical fire control application designed to assist in indirect fire operations.
Artillery Management	Armor-notepad	Tactical application for managing artillery operations and reconnaissance.
Artillery Management	D1-note	Artillery app for D-1 howitzer operations.
Ballistic Calculator	Strelok Pro	Advanced ballistic calculator.
Ballistic Calculator	Strelok+	Ballistic calculator.
Explosives Calculator	Calculator PR	Calculates explosive charges for various objects and materials.
Explosives Calculator	Engineer’s Directory	Explosives, grenades, and detonators calculator.
Field Logistics & Support	Leon	Tactical-technical characteristics of rifles, grenades, and special armaments.
Field Logistics & Support	ZMops SOFT	Military software inventory.
Field Logistics & Support	Control BK—Molot	Tracks ammunition inventory and usage for artillery systems.

Field Logistics & Support	Skrezhet	Reduces operator radio visibility, maximizes channel range, and ensures mobility and year-round weather resistance.
Mapping & Navigation	Dots	Software for encoding electronic maps (objects, routes, areas).
Mapping & Navigation	Z Map Viewer	Offline navigation app.
Mapping & Navigation	ZMops Maps	Navigation maps with integrations to various battle apps.
Mapping & Navigation	ZOV Maps	Mapping application designed for operational use.
Mapping & Navigation	ZOV Maps Demo	ZOV Maps Demo is a demonstration version of the ZOV Maps application.
Mapping & Navigation	Topogeodeziya SK-42	Field topography calculations using full or reduced coordinates within one zone or adjacent zones.
Medical & Training	VMedA Tactical Medicine	Guidance on first aid, managing injuries, and critical medical skills.
Meteorological Tools	Meteo	Meteorological data analysis for artillery, UAVs, and tactical planning.
Meteorological Tools	ZeVs—METEO	Provides meteorological data critical for artillery operations.
Meteorological Tools	ZeVs—METEO ALPHA	Provides meteorological data critical for artillery operations.
Tactical Communication	ZOV Chat	Communication app designed for devices running licensed versions of ZOV Maps, providing connectivity and chat functionality.
Tactical Communication	Groza	Tactical communication platform that integrates with drones, radios, and meteorological systems for data exchange, coordination, and fire adjustment.
Tactical Communication	Calculator STC	Quick calculations in radiocommunications, including visibility, signal attenuation, antenna gain, and coordinate conversions.
Tactical Communication	Loktar	Provides secure, low-visibility tactical communication, data transfer, mapping, and drone detection capabilities.
Tactical Communication	Malina	A Raspberry Pi-based system for integrating radios with networks, enabling secure tactical communication and device coordination.
UAV Management	Drone Detector	Identifies and tracks drones using frequency analysis.
UAV Management	DroneAlert	Monitors and alerts on detected radio signals, including those from drones.
UAV Management	Eye Lite	Tactical drone software system designed for reconnaissance, target identification, telemetry integration, and artillery fire adjustment.

UAV Management	FlyStat	UAV flight statistics.
UAV Management	Karlson3	Assists in drone operations, focusing on artillery fire correction, and providing tools like distance grids, offline maps, and directional calculations.
UAV Management	UavData	Organizes operational documentation and intelligence reporting during UAV missions.
UAV Management	ZeVs—BPLA	Supports drone operators in precise targeting and reconnaissance tasks.
UAV Management	Veterok	Part of the Veterok-ArtGruppa tactical software suite for reconnaissance and artillery units.
UAV Management	Veterok-T	Part of the Veterok-ArtGruppa tactical software suite for reconnaissance and artillery units.
UAV Management	Glaz 3	Tactical drone software system.
UAV Management	Glaz 4T	Tactical drone software system.
UAV Management	Glaz 2	Tactical drone software system.
UAV Management	Flight Log Avacha—Operator	Flight log app.
UAV Management	Trepet ID	Drone software designed for target identification, artillery fire adjustment, and ammunition drop assistance.

TABLE II: NETWORKED MILITARY APP CATEGORIES

Category	Number of Apps
Artillery Management	25
UAV Management	14
Mapping and Navigation	6
Tactical Communication	5
Field Logistics & Support	4
Meteorological Tools	3
Ballistic Calculator	2
Explosives Calculator	2
Medical & Training	1

The diverse array of app functionalities reflects a systematic effort to digitize and streamline military operations. It also underscores the strategic importance of technology in enabling effective battlefield management, situational awareness, and resource optimization, all critical components of contemporary military strategy.

4. EXAMPLES OF MILITARY APPS

To provide deeper insights into the functionality and technological dependencies of Russian military mobile applications, this section highlights notable examples from different categories identified during the study. These examples illustrate the diverse roles these apps play on the battlefield, from artillery management to reconnaissance and meteorological support.

Each app selected represents a critical component of military operations. We have highlighted each app's primary purpose, networked functionality, and reliance on external infrastructure. By analyzing these apps, we can better understand how they leverage global technological resources and what the implications of their usage are in the context of modern warfare.

The examples presented here serve to emphasize the operational sophistication and interconnectedness of these tools, shedding light on the broader ecosystem that enables their use in conflict scenarios.

ZeVs—METEO ALPHA

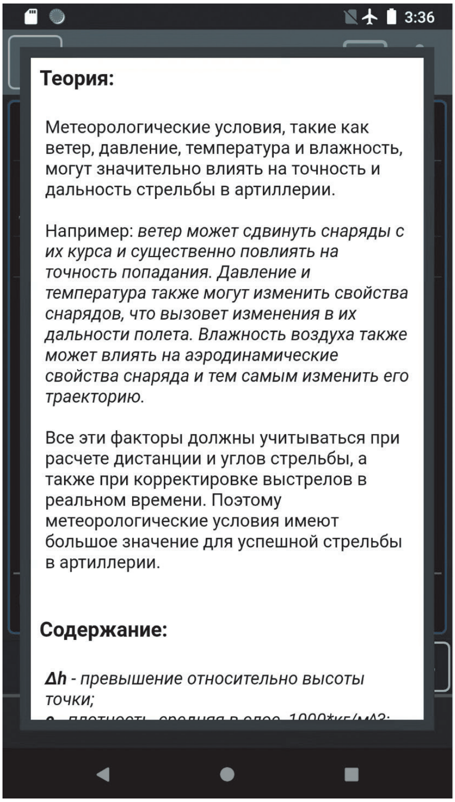
App name: ZeVs—METEO ALPHA
Package name: zevs.team.arta
Analyzed version: 2.1 (ZeVs—METEO 2.10 ALPHA.apk,
MD5: ab1ef838e792d4aeb4c6cd1551c39dab)
Category: Meteorological Tools

Description
ZeVs—METEO provides essential meteorological data to enhance artillery precision. It calculates deviations caused by atmospheric factors such as wind, air pressure, temperature, and humidity. The app generates detailed meteorological bulletins and facilitates real-time adjustments to firing angles and ranges based on environmental conditions. Additionally, ZeVs—METEO integrates seamlessly with other apps by the same developer group, as well as a broader ecosystem of military apps, enabling a cohesive operational environment for Russian forces.

Screen Translation

The functionality of ZeVs—METEO, as described in the mobile app's help section, is presented in Figure 1; translation follows.

FIGURE 1: HELP SCREEN OF ZeVs—METEO ALPHA, WITH TRANSLATION



Теория:

Метеорологические условия, такие как ветер, давление, температура и влажность, могут значительно влиять на точность и дальность стрельбы в артиллерии.

Например: *ветер может сдвинуть снаряды с их курса и существенно повлиять на точность попадания. Давление и температура также могут изменить свойства снарядов, что вызовет изменения в их дальности полета. Влажность воздуха также может влиять на аэродинамические свойства снаряда и тем самым изменить его траекторию.*

Все эти факторы должны учитываться при расчете дистанции и углов стрельбы, а также при корректировке выстрелов в реальном времени. Поэтому метеорологические условия имеют большое значение для успешной стрельбы в артиллерии.

Содержание:

Δh *- превышение относительно высоты точки;*
ρ *- плотность, средняя в слое, 1000*кг/м^3;*

Theory:

"Meteorological conditions, such as wind, pressure, temperature, and humidity, can significantly impact the accuracy and range of artillery fire.

"For example, wind can shift projectiles off course, greatly affecting targeting accuracy. Pressure and temperature can also alter projectile properties, resulting in changes to their flight range. Additionally, air humidity can affect the aerodynamic properties of the projectile, thereby altering its trajectory.

“All these factors must be taken into account when calculating shooting distances and angles, as well as when adjusting fire in real time. Therefore, meteorological conditions play a critical role in the success of artillery operations.”

Domains Used

The app communicates with a range of hosts, including mapping, topography, and cloud service providers:

- android.com (1 occurrence)
- autonavi.com (4 occurrences)
- chartbundle.com (1 occurrence)
- cloudmade.com (4 occurrences)
- google.cn (4 occurrences)
- google.com (6 occurrences)
- nationalmap.gov (1 occurrence)
- openptmap.org (1 occurrence)
- openseamap.org (1 occurrence)
- openstreetmap.nl (1 occurrence)
- openstreetmap.org (3 occurrences)
- opentopomap.org (3 occurrences)
- t.me (1 occurrence)
- tianditu.com (6 occurrences)
- wikimedia.org (1 occurrence)
- wmflabs.org (1 occurrence)
- zevstech.ru (1 occurrence)

Geographic Distribution

The app’s hosts are distributed across several countries, reflecting a reliance on international infrastructure:

- United States (US): 18 occurrences
- China (CN): 8 occurrences
- Germany (DE): 5 occurrences
- Netherlands (NL): 1 occurrence
- Russia (RU): 1 occurrence
- Antigua and Barbuda (AG): 1 occurrence

IP Ownership

The app relies on infrastructure from multiple global technology providers:

- Google: 9 occurrences
- Huawei: 10 occurrences
- Alibaba: 5 occurrences
- Amazon: 4 occurrences
- Deutsches Forschungsnetz: 4 occurrences
- Fastly: 3 occurrences
- Telegram: 1 occurrence
- TimeWeb: 1 occurrence

Analysis
The ZeVs—METEO ALPHA app not only demonstrates the reliance on a globally distributed network of services but also highlights its integration with other tools within the broader military app ecosystem. This interconnectedness amplifies the app's operational utility, allowing real-time data sharing and streamlined workflows in military contexts.

Symbolism in Naming
The letters Z and V, initially tactical markings on Russian military vehicles, have evolved into pro-war propaganda symbols. Z likely stands for "Zapad" "West" or "Za pobedu" "For victory", while V may signify "Vostok" "East" or "Victory." Widely adopted in Russian propaganda,[6] these symbols represent support for the invasion of Ukraine and are used to foster nationalism. Their inclusion in military apps like ZeVs—METEO ALPHA ties the tools to Russia's military identity and ideological objectives.

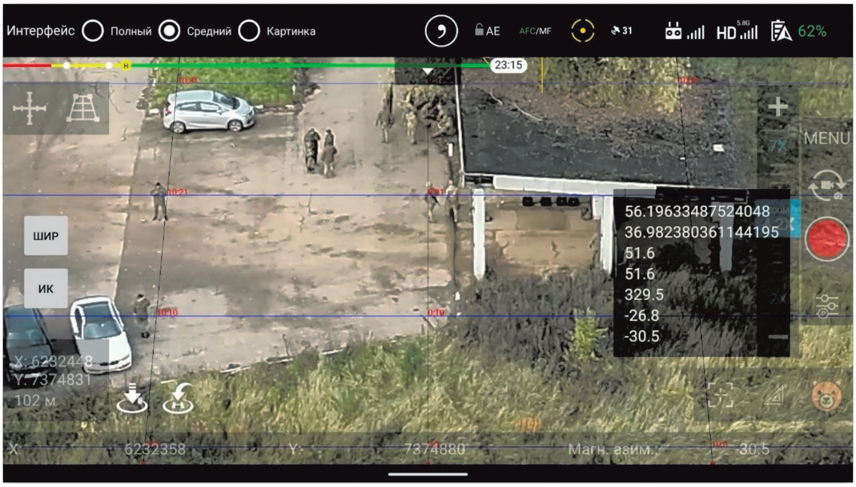
Karlson3

App name: Karlson3
Package name: ru.karlson
Version analyzed: 0.2.1 (Karlson3—0.2.1.apk,
MD5: 26e948f80909fdbdc0bee574efd3c7a4)
Category: UAV Management

Description
The Karlson (in Russian, Карлсон) app facilitates drone operations with a focus on artillery fire correction. It provides tools such as distance grids, offline maps, and directional calculations to support precision targeting. The app is compatible with various drone models from the Chinese tech company DJI, enhancing accuracy and operational efficiency in combat scenarios. The user interface of the Karlson mobile app is presented in Figure 2.

[6] Orysia Hrudka, "Why Have Z and V Become Russia's Pro-war Symbols?," Euromaidan Press, March 24, 2022, https://euromaidanpress.com/2022/03/24/why-do-z-and-v-become-russians-pro-war-symbols/.

FIGURE 2: KARLSON3 USER INTERFACE



Интерфейс ○ Полный ● Средний ○ Картинка

Domains Used

The app communicates with a wide range of domains, spanning drone management, mapping, cloud services, and general infrastructure:

- a9.com (1 occurrence)
- amap.com (11 occurrences)
- amazon.com (1 occurrence)
- amazonaws-china.com (1 occurrence)
- amazonaws.com (1 occurrence)
- apache.org (1 occurrence)
- autonavi.com (1 occurrence)
- biying.com (1 occurrence)
- chartbundle.com (1 occurrence)
- cloudmade.com (4 occurrences)
- creativecommons.org (1 occurrence)
- dji-flighthub.com (1 occurrence)
- dji.com (4 occurrences)
- dji.net (1 occurrence)
- djistatic.com (1 occurrence)
- georss.org (1 occurrence)
- github.com (1 occurrence)
- githubusercontent.com (1 occurrence)
- godaddy.com (3 occurrences)

- google.com (3 occurrences)
- gov.cn (1 occurrence)
- ionicons.com (1 occurrence)
- mapbox.com (4 occurrences)
- maptiler.com (1 occurrence)
- nationalmap.gov (1 occurrence)
- openptmap.org (1 occurrence)
- openseamap.org (1 occurrence)
- openstreetmap.org (1 occurrence)
- opentopomap.org (3 occurrences)
- thunderforest.com (3 occurrences)
- virtualearth.net (1 occurrence)
- wikimedia.org (1 occurrence)
- wmflabs.org (1 occurrence)
- xmlpull.org (1 occurrence)
- zetetic.net (1 occurrence)

Geographic Distribution

The app’s infrastructure spans multiple countries, reflecting its reliance on global technology resources:

- United States (US): 44 occurrences
- China (CN): 19 occurrences
- Germany (DE): 9 occurrences
- Netherlands (NL): 1 occurrence
- United Kingdom (GB): 1 occurrence

IP Ownership

The app leverages services from major technology providers and networks:

- Alibaba: 12 occurrences
- Amazon: 11 occurrences
- Cloudflare: 7 occurrences
- Hetzner: 6 occurrences
- Deutsches Forschungsnetz: 6 occurrences
- Microsoft Corporation: 3 occurrences
- GitHub: 1 occurrence
- Google: 2 occurrences

Analysis

Karlson3 underscores the integration of drone-specific functionalities with tools for precision artillery targeting. Its reliance on a mix of Western and Chinese infrastructure highlights the global interconnectedness of the military-app ecosystem. Key dependencies include DJI's ecosystem for drone operations and mapping platforms like Mapbox and CloudMade for geospatial intelligence. The app's use of multiple mapping and cloud services mirrors its focus on operational accuracy and redundancy.

Symbolism in Naming

The name Karlson in the app Karlson3 references the character Karlsson-on-the-Roof (in Russian, Карлсон, который живет на крыше), a children's book character created by Swedish author Astrid Lindgren. Karlsson is a mischievous man with a propeller on his back, allowing him to fly—arguably a symbolic nod to the app's focus on drone operations.

Veterok

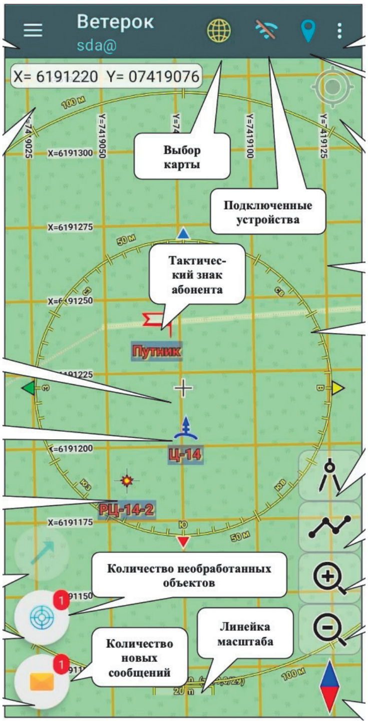
App name: Veterok
Package name: ru.niissu.veterok
Version analyzed: 1.16.2 (Ветерок—1.16.2.apk,
MD5: 369939097892f0d57f8e9ae24ba398a0)
Category: Artillery Management

Description

Veterok is part of the Veterok-ArtGruppa (in Russian, Ветерок-АртГруппа,) complex, a tactical software suite designed for reconnaissance and artillery units. The app supports object detection, data transmission, artillery fire adjustment, and geographical calculations. Its focus on reconnaissance-strike automation[7] makes it a critical tool in integrating intelligence and artillery fire operations. The tactical screen of the Veterok-ArtGruppa mobile app is presented in Figure 3.

[7] The term "reconnaissance-strike complex" refers to an integrated military system that combines real-time intelligence gathering with precision-strike capabilities to engage high-value targets efficiently. This concept, developed by the Soviet Union and later revived by Russia, utilizes advanced surveillance, automated command and control, and long-range precision weapons to detect and destroy targets swiftly.

FIGURE 3: VETEROK MAIN SCREEN AS ILLUSTRATED IN A USER MANUAL



Domains Used

The app connects to various domains for mapping, data integration, and cloud services:

- 2gis.com (1 occurrence)
- arcgisonline.com (1 occurrence)
- chartbundle.com (1 occurrence)
- cloudmade.com (4 occurrences)
- google.com (2 occurrences)
- hereapi.com (1 occurrence)
- nakarte.me (1 occurrence)
- nationalmap.gov (1 occurrence)
- opengis.net (1 occurrence)
- openptmap.org (1 occurrence)
- openseamap.org (1 occurrence)
- openstreetmap.nl (1 occurrence)
- openstreetmap.org (4 occurrences)

- opentopomap.org (4 occurrences)
- telegram.me (1 occurrence)
- 12andex12ia.org (1 occurrence)
- wmflabs.org (1 occurrence)
- xmlpull.org (1 occurrence)
- 12andex.net (3 occurrences)

Geographic Distribution

The infrastructure supporting the app spans multiple countries, emphasizing global dependencies:

- United States (US): 18 occurrences
- Germany (DE): 7 occurrences
- Russia (RU): 3 occurrences
- Netherlands (NL): 1 occurrence
- Antigua and Barbuda (AG): 1 occurrence

IP Ownership

The app relies on a range of technology providers for its functionality:

- Amazon: 6 occurrences
- Cloudflare: 1 occurrence
- Google: 2 occurrences
- Deutsches Forschungsnetz: 4 occurrences
- Fastly: 4 occurrences
- Hetzner: 1 occurrence
- NetCup: 2 occurrences
- Telegram: 1 occurrence
- Wikimedia: 1 occurrence
- Yandex: 3 occurrences

Analysis

Veterok exemplifies the integration of reconnaissance and artillery-fire automation through its comprehensive tactical features. The app's reliance on mapping and geospatial services (e.g., OpenStreetMap, ArcGIS, HERE) underscores its focus on precision and situational awareness. With a mix of global and Russian infrastructure, Veterok demonstrates a dual dependency on Western technology and localized Russian resources, reflecting the complexity of modern military software ecosystems. Its integration with communication platforms like Telegram further enhances its utility in real-time battlefield scenarios.

Symbolism in Naming

The names Veterok ("light breeze") and ArtGruppa ("artillery group") reflect their military roles. Veterok symbolizes agility and real-time reconnaissance, aligning with its use in dynamic operations, while ArtGruppa directly references artillery coordination, emphasizing its tactical purpose in team-based fire adjustments. Both names are practical and resonate with their operational contexts.

5. ANALYSIS RESULTS

The analysis of the selected military mobile applications revealed extensive use of online services to support their operational functions. Across the functionality, configuration, and resources of these apps, a total of 1,594 network addresses (DNS hostnames and IP addresses) were identified. To ensure accurate representation and avoid duplication, each service and IP address was accounted for only once. This resulted in a refined dataset of 323 distinct hostnames and 387 corresponding distinct IP addresses that was subjected to further analysis.

Each of these distinct data elements was examined to determine its geographical location and network ownership. The GeoIP analysis provided insights into the global distribution of resources utilized by the apps, while the ownership information revealed the entities responsible for hosting these services. The findings highlight a significant reliance on infrastructure provided by international cloud service providers, VPS platforms, and cybersecurity services.

The detailed breakdown of these results, including ownership distribution and geographical spread, is provided in the following sections.

A. Identified Domains

The examination of 323 hostnames embedded in the source code of Russian military apps revealed a distribution across 204 distinct level-2 domains. Below is a detailed analysis of the top 20 domains:

1. cloudmade.com (132 occurrences)

CloudMade.com was a mapping and navigation platform that leveraged OpenStreetMap data and later transitioned to AI-driven services for the automotive industry.

2. openstreetmap.org (110 occurrences)

OpenStreetMap.org is a collaborative, open-source mapping platform providing free, editable geospatial data used for various applications worldwide.

3. opentopomap.org (105 occurrences)
OpenTopoMap.org is an open-source mapping platform offering topographic maps generated from OpenStreetMap data, tailored for outdoor and geographical use.

4. mapbox.com (103 occurrences)
Mapbox.com is a platform providing customizable mapping and geospatial tools, including APIs and SDKs, for developers to integrate location-based features into applications.

5. google.com (79 occurrences)
Google's popularity here reflects a reliance on its ecosystem for APIs, backend data handling, and other infrastructure services.

6. wmflabs.org (72 occurrences)
Wmflabs.org is a hosting platform for Wikimedia Foundation projects, supporting development, testing, and tools related to Wikimedia's open knowledge initiatives.

7. amap.com (36 occurrences)
Amap.com is a Chinese mapping and navigation service, also known as AutoNavi, providing real-time traffic, location-based services, and geospatial data.

8. openptmap.org (34 occurrences)
OpenPtMap.org visualizes public transport networks using OpenStreetMap data, offering a clear overview of transit routes and infrastructure.

9. nationalmap.gov (34 occurrences)
NationalMap.gov is a US government platform providing geospatial data, including topographic maps and environmental datasets, for public and professional use.

10. chartbundle.com (34 occurrences)
Chartbundle.com was a hobbyist website offering digital aviation charts for flight planning (now discontinued), with its source code available on GitHub.

11. openstreetmap.nl (33 occurrences)
OpenStreetMap.nl is the Dutch OpenStreetMap community hub, providing resources and tools for collaborative map editing in the Netherlands.

12. openseamap.org (33 occurrences)
OpenSeaMap is a free, worldwide nautical chart project that enhances OpenStreetMap with maritime information, including sea marks, harbors, and water depths, to support navigation and marine activities.

13. github.com (32 occurrences)
GitHub.com is a platform for version control and collaborative software development, enabling users to host, share, and manage code repositories.

14. android.com (30 occurrences)
Android.com is the official website for Google's Android operating system, offering resources for users, developers, and device manufacturers.

15. thunderforest.com (27 occurrences)
Thunderforest.com provides customizable map styles and APIs for outdoor activities, built on OpenStreetMap data, catering to developers and enthusiasts.

16. xmlpull.org (25 occurrences)
Xmlpull.org is a resource for the XML Pull Parser API, offering lightweight, efficient tools for XML parsing in Java-based applications.

17. tilestream.net (19 occurrences)
Tilestream.net is a platform for hosting and serving custom map tiles, enabling developers to create and manage personalized map visualizations.

18. firebaseio.com (19 occurrences)
Firebaseio.com is a domain used by Google Firebase to provide backend services like real-time databases, authentication, and cloud functions for applications.

19. googlesyndication.com (16 occurrences)
Googlesyndication.com is a domain used by Google for delivering ads, dynamic content, and resources through its advertising and content platforms.

20. wikimedia.org (15 occurrences)
Wikimedia.org is the official domain of the Wikimedia Foundation, hosting projects like Wikipedia and providing free, open-access knowledge and resources.

The data reveals a heavy reliance of Russian military apps on open-source, commercial, and global cloud platforms for critical functionalities like mapping, navigation, and backend operations. Open-source tools such as openstreetmap.org and commercial platforms like mapbox.com provide customizable geospatial data, while global cloud providers, such as Google (google.com, firebaseio.com), enable real-time data handling and app distribution. Public resources like nationalmap.gov and niche platforms like openseamap.org demonstrate how freely available data and specialized tools are repurposed for military objectives.

This domain-level analysis illustrates how military apps leverage a mix of global, open-source, and commercial resources to achieve military operational efficiency. The widespread use of publicly available platforms raises critical questions about their unintended use in conflict scenarios, further emphasizing the ethical and legal complexities of technology in modern warfare. See the list of identified domain names in Table III.

TABLE III: TOP 20 DOMAIN NAME OCCURRENCES IN MILITARY APPS

Domain Name	Number of Occurrences
cloudmade.com	132
openstreetmap.org	110
opentopomap.org	105
mapbox.com	103
google.com	79
wmflabs.org	72
amap.com	36
openptmap.org	34
nationalmap.gov	34
chartbundle.com	34
openstreetmap.nl	33
openseamap.org	33
github.com	32
android.com	30
thunderforest.com	27
xmlpull.org	25
tilestream.net	19
firebaseio.com	19
googlesyndication.com	16
wikimedia.org	15

B. Geographic Distribution

The geographical distribution of the 387 distinct IP addresses supporting the networked military apps underscores a heavy reliance on infrastructure located in the United States, with 283 IPs (73%) linked to US-based services. This dominance reflects the widespread use of global cloud service providers headquartered in the US.

Other significant contributors include China (30 IPs, 8%) and Germany (24 IPs, 6%), suggesting secondary hubs for hosting and infrastructure. Notably, Russia (17 IPs, 5%) ranks fourth, representing locally managed or proximate services supporting the military apps.

European nations such as France (6 IPs), Switzerland (4 IPs), the Netherlands (3 IPs), Finland (3 IPs), Ireland (2 IPs), the United Kingdom (2 IPs), Bulgaria (1 IP), Romania (1 IP), and Italy (1 IP) collectively account for 12% of the distribution. This reflects the involvement of various hosting and infrastructure providers across Europe.

A smaller number of IPs were distributed across other regions, including Singapore (3 IPs), New Zealand (1 IP), Japan (1 IP), Australia (1 IP), and Antigua and Barbuda (2 IPs). The presence of Ukraine (1 IP) is notable but minimal.

This distribution highlights the global reach and dependence of these military apps on foreign infrastructure, with US and European services playing a particularly critical role. The geographical distribution of identified IP addresses is presented in Table IV and in Figure 4.

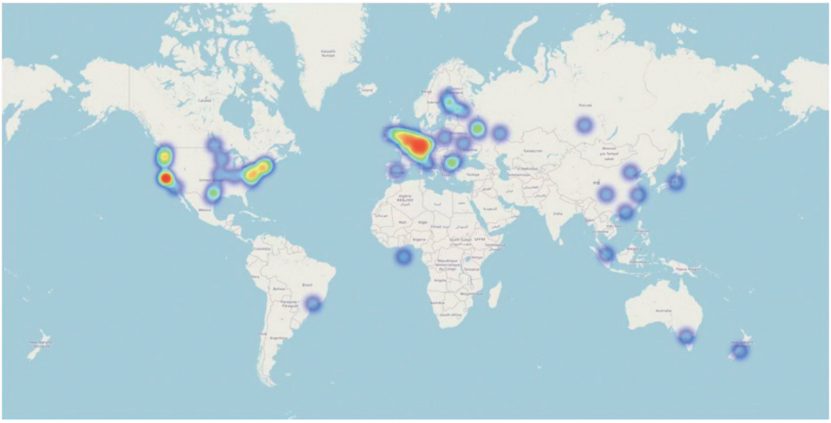
TABLE IV: COUNTRY CODES OF AUTONOMOUS SYSTEMS FOR IDENTIFIED IP ADDRESSES

ASN Country Code	Number of Hosts
US	283
CN	30
DE	24
RU	17
FR	6
CH	4
SG	3
NL	3
FI	3
IE	2
GB	2
AG	2
UA	1
RO	1
NZ	1
JP	1
IT	1
BG	1
AU	1

C. IP Address Ownership

The analysis of the 387 distinct IP addresses revealed a clear pattern of ownership concentrated among major technology and infrastructure providers. This distribution highlights the heavy reliance of Russian military mobile applications on well-established international platforms.

FIGURE 4: HEAT MAP OF IP ADDRESSES' GEOGRAPHIC DISTRIBUTION



1. Dominant Cloud Service Providers

- Amazon Web Services (127 IPs, 33%): AWS accounts for over a third of the IPs analyzed, reflecting its dominant position in global cloud hosting and backend support services.
- Google Cloud Platform (53 IPs, 14%): GCP is the second-largest provider, underscoring its widespread use for storage, APIs, and data processing.
- Microsoft (5 IPs) and DigitalOcean (4 IPs) also play notable roles, albeit on a smaller scale.

2. Content Delivery and Network Security

- Cloudflare (30 IPs, 8%): Cloudflare is known for its Distributed Denial of Service (DDoS) protection and Content Delivery Network (CDN) services. Its infrastructure is a critical enabler of secure communication for respective apps.
- Fastly (14 IPs) and Akamai (4 IPs) contribute additional content delivery capabilities.
- Sucuri (4 IPs) provides security and monitoring solutions.

3. Chinese Technology Giants

- Alibaba Cloud (26 IPs, 6%) and Huawei (10 IPs, 3%) reflect the significant involvement of Chinese infrastructure providers, with Alibaba serving both hosting and database needs.

4. European Hosting Providers

- Hetzner (18 IPs, 5%) and OVH (3 IPs) represent key European hosting platforms, often favored for their affordability and reliability.

5. Russian-Owned Services

- Yandex (8 IPs) and SprintHost (3 IPs) are Russian-owned, reflecting local services utilized in some cases to ensure proximity and compliance with Russian regulations.

The reliance on US-based platforms such as Amazon, Google, and Cloudflare underscores the paradox of Western infrastructure supporting adversarial military apps. Chinese providers like Alibaba and Huawei play a smaller but significant role, highlighting a secondary dependency on non-Western platforms. European providers such as Hetzner and OVH are also part of the ecosystem, further diversifying the technological dependencies of these apps.

D. Developer Organization and Structure

Russian military app development is largely driven by civilian developers, not official military programmers. While a few applications have reached the final stages of official adoption, requiring companies to restructure as joint stock companies with state ownership, most remain relatively independent grassroots projects.

Developers come from various backgrounds, including:

- Volunteers ideologically motivated to contribute to military tech.
- Private sector employees developing tools as part of a company’s bid to secure government contracts.
- Academic or scientific institutions providing research and development capacity.
- Crowdfunded or unofficially compensated groups of individuals.

Most developers label their projects as “инициативные разработки” “initiative-based development” or “перспективные разработки” “promising developments”, indicating that these apps are not yet officially adopted but are intended for eventual military integration.

Nonetheless, these military apps are already widely used in operations. The most successful ones have thousands to tens of thousands of active users, depending on their complexity and purpose. Their development teams and sponsoring organizations

conduct regular training sessions for military personnel, including in active operational zones, and oversee testing exercises at military facilities. Some apps have even been formally documented in official military textbooks, detailing their functionality and usage in combat scenarios, and integrated into military institutes' coursework.
The military mobile apps ecosystem represents a hybrid private-public partnership, where software development is loosely coordinated but strictly aligns with official military requirements.

E. Rationale Behind Cloud Selection

It is reasonable to conclude that Russian military apps use Western cloud services and APIs not out of deliberate preference, but because they are more accessible, cost-effective, and reliable than domestic alternatives—if such alternatives even exist. These choices are shaped by infrastructure convenience, data availability, and the advantages of mature developer ecosystems, rather than strategic selection. Notably, four leading Russian cloud providers—Rostelecom, Cloud.ru, Selectel, and MTS—are conspicuously absent from this ecosystem. A fifth, Yandex Cloud,[8] is minimally represented, hosting only eight IP addresses identified in this study.

As this study demonstrates, cloud service providers play a critical role in military app development by hosting or securing essential services, such as meteorological forecasts and geospatial intelligence. This explains why Western clouds predominate at the infrastructure (e.g., IP address) level in the observed results.

Russia lacks high-quality meteorological and geospatial data, making Western sources indispensable.[9] Accurate weather forecasting requires supercomputing resources, which Russia struggles to maintain.[10] And while Russia has domestic mapping services, they do not provide the high-precision datasets needed for targeting and artillery support.[11] Particularly, Russia's reliance on OpenStreetMap and commercial geospatial APIs suggests that domestic mapping lacks the necessary resolution and detail.

Google Cloud and Firebase are widely used because they are practically free via the Free Tier GCP offering and the no-cost Firebase Spark plan, require no official developer accounts, have extensive documentation, and are broadly popularized by online training programs and video tutorials.

[8] "Бизнес сгущает облака," Коммерсант, August 8, 2024, https://www.kommersant.ru/doc/6879530.

[9] Reade Levinson, "Russia Receives Western Weather Data That Some Fear Could Aid Attack Planning," Reuters, March 22, 2022, https://www.reuters.com/article/world/russia-receives-western-weather-data-that-some-fear-could-aid-attack-planning-idUSKCN2LJ0U5/.

[10] "Суперкомпьютеры, в том числе задействованные в прогнозировании погоды в России могут продолжить работу в течении трех лет," Метеожурнал, April 28, 2022, https://meteojurnal.ru/superkompyutery-v-tom-chisle-zadejstvovannye-v-prognozirovanii-pogody-v-rossii-mogut-prodolzhit-rabotu-v-techenii-treh-let/.

[11] Michael Peck, "Why Russian Space Satellites Are Failing in the Ukraine War," Popular Mechanics, March 29, 2023, https://www.popularmechanics.com/military/a43444628/why-russian-satellites-are-failing-in-ukraine/.

Unlike Apple's App Store,[12] Google's Play Market, with its less restrictive policies, enables easy sideloading and unofficial distribution channels,[13] making Android the preferred platform for Russian military apps.

F. Policy Prescriptions

To mitigate the exploitation of Western technology, targeted restrictions must be implemented at multiple levels. The key concern here is the high-threat areas, where the development or use of military apps takes place. Establishing criteria or identifying high-threat areas falls beyond this study's scope and involves international law and IHL considerations. For this study, we set the scope of high-threat areas to states that sponsor or conduct illegal wars of aggression and the territories they occupy or annex.

Cloud Service Providers (AWS, Cloudflare, etc.)

- Enhance verification processes for user origins to limit access from sanctioned and high-threat areas.
- Implement geofencing for high-threat areas using user IP addresses and more sophisticated geolocation technologies.
- Proactively identify military usage in high-threat areas in the way current measures against universally illegal activities, such as child exploitation content, are implemented.

SDK and Developer Ecosystem Providers (Google)

- Limit access to development tools (SDKs, APIs, supporting cloud services, distribution channels, etc.) for users from high-threat areas like Apple did for enterprise developers from Russia in February 2025.[14]
- Restrict access to sensitive APIs, including meteorological and geospatial services, exclusively to applications distributed through official app stores, thereby preventing the unauthorized sideloading of critical software.

Critical Applications and Services (Meteorology, Cartography, etc.)

- Extend abuse policies to include military applications developed and used in high-risk areas.
- Establish reporting mechanisms to identify and eliminate access to services that are used for military purposes in the high-threat areas.

[12] "Building a Trusted Ecosystem for Millions of Apps: A Threat Analysis of Sideloading," Apple, October, 2021, https://www.apple.com/privacy/docs/Building_a_Trusted_Ecosystem_for_Millions_of_Apps_A_Threat_Analysis_of_Sideloading.pdf.

[13] "Alternative Distribution Options," Google Android Developers, last updated April 16, 2020, https://developer.android.com/distribute/marketing-tools/alternative-distribution.

[14] "Apple закрыла россиянам доступ к платформе разработки бизнес-приложений," РБК, February 24, 2025, https://www.rbc.ru/technology_and_media/24/02/2025/67b9be389a79470a2de2be8a.

Regulatory Actions

- Establish legal accountability for cloud providers that offer services to unverified or malicious users in high-risk areas.
- Create official reporting channels for national CERT teams to flag military applications from high-risk areas for review and possible removal.
- Prohibit military software developers originating from high-risk areas from employment in Western companies or obtaining residency in Western countries.
- Launch national bug bounty programs to crowdsource intelligence on apps used in military aggression.

6. CONCLUSION

The study reveals the rise of a grassroots “people’s military-industrial complex” in Russia, leveraging open-source and Western technological resources to create military Android applications. These apps reflect strong cultural cohesion and operational innovation, becoming highly effective tools in wartime. Their success is facilitated by the openness of the Google Android ecosystem and the unrestricted access to global cloud services, which collectively enable their rapid development, scale, and distribution.

The exploitation of these platforms raises serious ethical and strategic concerns. Major cloud providers and open ecosystems inadvertently support this ecosystem by failing to implement controls that could restrict access by the states prosecuting illegal war or prevent the weaponization of their resources. This reliance on global infrastructure highlights a critical gap in accountability and governance in the technology sector during conflict.

To counter this exploitation, access restrictions and regulatory controls are crucial. Cloud service providers must implement regional access controls, monitor resource usage, and restrict applications exploited for unlawful military operations. Developer ecosystems must deny access to programming toolkits and critical services for software developers engaged in wars of aggression. Providers of essential services, such as geospatial intelligence and meteorological data, must ensure their platforms are used solely for legitimate purposes. Regulators should enforce compliance by incentivizing adherence to these measures while penalizing violators.

Limiting access to global infrastructure would compel Russian developers to depend on less capable and reliable domestic resources and attempt to circumvent network access controls by additional means such as VPN, undermining their operational effectiveness. These measures are vital for curtailing the misuse of global technology in illegal warfare while preserving the ethical integrity of open and commercial platforms.

CyCon 2025: The Next Step
17th International Conference on Cyber Conflict
C. Kwan, N. Gratzer, K. Podiņš, M. Tolppa (Eds.)
2025 © NATO CCDCOE Publications, Tallinn

Permission to make digital or hard copies of this publication for internal use within NATO and for personal or educational use when for non-profit or non-commercial purposes is granted provided that copies bear this notice and a full citation on the first page. Any other reproduction or transmission requires prior written permission by NATO CCDCOE.

Towards Technology-Based Regulation of China-Made IoT Surveillance IP Cameras: The Case Study of Australia

Ausma Bernot*
Lecturer
School of Criminology
and Criminal Justice
Griffith University
Gold Coast, QLD, Australia
a.bernot@griffith.edu.au

M. Arif Khan
Senior Lecturer
School of Computing and Mathematics
Charles Sturt University
Wagga Wagga, NSW, Australia

Khurram Shahzad
Research Officer
School of Computing and Mathematics
Charles Sturt University
Wagga Wagga, NSW, Australia

Mert Karakaya
Senior Research Engineer
IPVM
Bethlehem, PA, United States

Conor Healy
Director of Government Relations
IPVM
Bethlehem, PA, United States

Abstract: Internet of Things (IoT) devices are currently under-regulated in Australia and in most countries/regions globally. Installing IoT devices on private property can cause security issues if an individual, business or institution becomes a target. When IoT devices are put in critical locations, such as federal or state government buildings, the likelihood of that location being a surveillance target rises. This article examines a recent case study on national security concerns related to China-made IoT Internet Protocol (IP) cameras in Australia, which were removed without any publicly disclosed technical tests. Our two-stage interdisciplinary research paper took the following steps: first, we used the Common Vulnerability Scoring System (CVSS) framework to assess the vulnerabilities of three IoT IP camera providers—Hikvision,

* This work was supported by the Griffith Criminology Institute, Griffith University, and the AI Cyberfutures Institute, Charles Sturt University. IPVM contributed in kind by providing IoT IP camera access.

Dahua, and Avigilon—that have been installed on federal government buildings. We found vulnerabilities in all three systems; however, Avigilon devices did not have any high or critical vulnerabilities, unlike Hikvision and Dahua. We then compared our findings to Australia's existing IoT device regulation frameworks. The present Australian regulations overlap and do not adequately address the existing cyber vulnerabilities. Instead, the security frameworks, recommendations, and legislation focus on organizational cyber hygiene and environmental security. Technical cyber security frameworks are classified and currently provided only on demand to certain federal government departments, leaving industry actors, state governments, and consumers without guidance. The Australian experience shows that uniform and mandatory cyber security requirements could improve the benchmark of IoT security while also benefiting consumers. As the European Union implements the Cyber Resilience Act to regulate software and hardware products with digital components, the effectiveness of cyber security enhancements for IoT devices will be tested.

Keywords: *cyber security, cyber supply chain, cyber vulnerabilities, IoT security, IoT regulation*

1. INTRODUCTION

The number of Internet of Things (IoT) devices worldwide has been estimated at between 21.5 billion (Sinha 2024) and 75 billion (Alam 2018) in 2025,[1] and this figure is expected to grow. In Australia, there were an estimated 371 million IoT devices as of 2024 (IoT Security Foundation 2021). A 2023 Telsyte Australian Smart Home Market Study survey (n=1,036) indicated a high level of internet-connected technologies, with the average household containing 23.8 IoT devices (16.1 non-smart devices, 7.1 smart home devices, and 0.5 provisioning devices). Despite high usage rates encouraged by the maturation of 5G and scaling of IoT device manufacture, security researchers indicate IoT devices remain vulnerable.

The major vulnerabilities in IoT devices include weak, guessable, or hard-coded passwords; insecure network services; insecure interfaces within ecosystems; lack of secure mechanisms for updates; the use of outdated or insecure components; insufficient protection of user privacy; insecure transfer and storage of data; poor device management practices; insecure default settings; and lack of physical security measures (Open Web Application Security Project n.d.). While exact statistics vary,

[1] Calculating the exact number of IoT devices globally is challenging due to varying definitions, rapid growth, decentralized manufacturing, shadow IoT, inconsistent tracking mechanisms, proprietary networks, fragmented standards, variable lifespans, geographical disparities, and the existence of illicit or undocumented devices.

cameras make up between 1% and 5% of all IoT devices (CUJOAI 2021; Palo Alto Networks 2021). IoT Internet Protocol (IP) cameras are, however, the largest segment of exposed IoT devices discoverable by Shodan, a search engine for Internet-connected devices, like webcams and routers (Siwakoti et al. 2023). Palo Alto Networks' Unit 42 (2021) analysis of 1.2 million enterprise and healthcare IoT devices similarly found that security cameras make up 5% of enterprise IoT devices but account for 33% of all security issues.

Deploying IoT IP cameras in homes can lead to data security issues, especially if a person or location is targeted. Criminal cases reported include hacking, recording and sales of child sexual abuse materials, and coercive control by an abusive partner to monitor the location and behavior of their victims (Brown, Harkin, and Tanczer 2024). IoT IP cameras installed in critical locations, such as government buildings, are at a higher risk of becoming targets for surveillance or hacking and are a national security concern. The security risks are not hypothetical. Compromised cameras can be exploited to capture, monitor, and sell inappropriate content, as well as facilitate burglary, stalking, and state-level crimes, including politically motivated offences (Blythe and Johnson 2021). During Nancy Pelosi's visit to Taiwan in 2023, IoT poster boards were hacked to display messages critical of her visit (Chen 2022).

IoT system targeting has the capacity to compromise not only the device but also the network to which the IoT system is connected. The threat of network compromise is exemplified by the large-scale 2016 Mirai botnet case, which compromised hundreds of thousands of IoT devices globally, including IP cameras, to launch DDoS attacks through host networks accessed via these hacked devices (Antonakakis et al. 2017). The release of Mirai's source code was a pivotal moment in IoT (in)security, leading to the emergence of numerous variants and inspiring other botnets, such as Okiru, Satori, Masuta, and PureMasuta, which continue to exploit vulnerabilities in IoT devices, including IP cameras (Guo et al. 2025). Focusing on IoT cyber security can help manufacturers mitigate some of these risks.

Chinese IoT IP camera manufacturers Dahua and Hikvision, the largest global providers, have faced significant scrutiny due to poor security standards and national security concerns. In 2021, independent researchers using Shodan data detected 6.3 million networks globally outside China, including over 41,000 Hikvision and 18,000 Dahua cameras in Australia (Migliano and Woodhams 2021). These cameras have been restricted in several countries, including the United States, the United Kingdom, Denmark, the Netherlands, and Australia (see Table I). Some countries, like Lithuania, have conducted cyber security testing and risk evaluations without firmly recommending replacements (NKSC 2020). Australia's ban on China-made cameras in federal government buildings was influenced by similar bans in allied countries and the prevalence of China-made IoT IP cameras in these buildings (Bernot and Smith 2023). The survey of Dahua and Hikvision cameras on federal buildings in Australia identified over 1,000 cameras, which were replaced by Avigilon.

TABLE I: A SUMMARY OF DAHUA AND HIKVISION RESTRICTIONS GLOBALLY

Country	Restriction	Ban	Method of Determining Cyber Risk
Australia	•		A survey of the IP cameras installed on federal government buildings led to their removal from federal government buildings.
Denmark	•		The Danish Centre for Cyber Security warned public entities that surveillance equipment purchased from entities covered in the US Entity List is a big security risk. The Danish Road Directorate accepted the advice and opted to replace roadside equipment (Lauritzen 2024).
India	•		In April 2024, India's Ministry of Electronics and Information Technology expanded the national Electronics and Information Technology Goods Order to include IP cameras to pass essential security requirements in India-based security laboratories before making the sales of those products available in India. At the time of writing, the requirements are set to begin in April 2025. The requirements include five categories of tests: physical security, access controls, network security, software security, and penetration testing.
Netherlands		•	The Netherlands decided to phase out China-made IP cameras due to concerns over human rights and espionage (NL Times 2024). The Association of Netherlands Municipalities flagged manufacturers for human rights violations (NL Times 2024). Additionally, the General Intelligence and Security Service of the Netherlands raised concerns about espionage, citing existing cyber attacks and espionage from China (AIVD 2022).
Taiwan		•	Due to national security concerns, in 2022, Taiwan's Ministry of National Defense and other government bodies have prohibited the procurement and use of Chinese information and communications products, including surveillance cameras from Hikvision and Dahua (Chen 2022).
United Kingdom		•	After the Government Security Group examined the existing and potential security risks of security cameras from China-based companies subject to the National Intelligence Law, the devices were pre-emptively removed from "sensitive sites" in 2022 (Federal Bureau of Investigation, Cyber Security and Infrastructure Security Agency, and National Security Agency 2022).
United States		•	In 2021, the Federal Communications Commission (FCC) banned Hikvision and Dahua equipment due to national security concerns over potential espionage, significantly restricting their import, sale, and use in the US, particularly in government and critical infrastructure sectors (FCC 2022).

IoT devices also remain under-regulated in Australia and in most countries/regions globally, thus posing a significant cyber security risk. The European Union's recent Cyber Resilience Act 2024 is one of a few global regulatory exceptions. When fully implemented, it will require manufacturers and distributors to report vulnerabilities and attend to security updates throughout the lifecycle of IoT products. Each country has the right to determine the best way to deploy security technologies in the interest of national security. However, national bans or individual restrictions are limited, as they can only target specific manufacturers, products, or groups of products within a particular country. Additionally, due to the complexity of the global supply chain, these restrictions can be and have been bypassed through practices such as white-labeling or redirecting partial production to different countries, thereby obfuscating the place of manufacture (Scanlan and Healy 2024).

We contend that existing cyber security methodologies and policy reviews can be more effective than ad hoc supplier bans in enhancing the overall standard of IoT security. Therefore, this article seeks to understand how technical IP camera security vulnerabilities interact with national and cyber security regulations in Australia. To this end, we use an interdisciplinary methodology that combines a technical methodology with a regulatory evaluation. Namely, we first deploy the Common Vulnerability Scoring System (CVSS) framework to test three IoT IP camera providers that were part of a politicized 2023 debate on the appropriate IoT IP cameras used. Second, we draw on the technical findings to determine whether the current Australian regulatory frameworks sufficiently cover the vulnerabilities found. Finally, the article considers lessons learned from Australia's regulatory approach to Chinese IP cameras.

2.TECHNICAL RESULTS: THE CVSS AND HISTORICAL RECORDS

The first part of the methodology involved a cyber security evaluation of Hikvision, Dahua, and a comparison manufacturer, Avigilon. Hikvision and Dahua were chosen because they are the two leading global suppliers of IoT IP cameras; Avigilon, a Canada-headquartered manufacturer, was selected for comparison because the company's cameras replaced Hikvision and Dahua on federal buildings in Australia, as confirmed by the first author during six site visits in February 2023.

A. CVSS Testing Framework

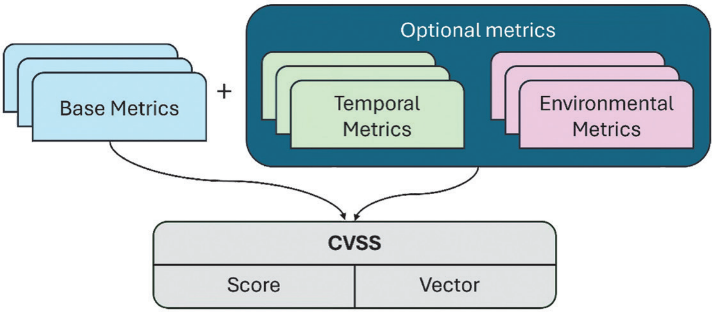
We chose to use the CVSS and the National Institute of Standards and Technology (NIST) guidelines, as they provide a well-established framework for delivering qualitative measures of vulnerabilities.[2] CVSS is a validated tool for vulnerability testing and remains a widely used standardized method for conducting comprehensive

[2] For a complete overview of CVSS, see https:// www.first.org/cvss/.

reviews of cyber security metrics and prioritizing remediation efforts. To prioritize remediation efforts, CVSS assigns numerical severity scores to vulnerabilities based on their potential impact. These scores facilitate comparisons across different manufacturers and products. For instance, the NIST uses CVSS to provide standardized vulnerability scores for federal agencies (Mell, Scarfone, and Romanosky 2007). Likewise, in the EU, the European Union Agency for Cybersecurity (ENISA) evaluates and communicates the severity of vulnerabilities across EU member states, following principles outlined in the CVSS methodology (ENISA 2024).

CVSS is linked to the Common Vulnerabilities and Exploitations (CVE) program, which identifies, defines, and catalogs publicly disclosed cyber security vulnerabilities (MITRE 2021; see Figure 1). CVSS rates these CVEs, with higher scores (closer to 10) indicating more severe vulnerabilities. Anyone can report a CVE by following a structured reporting process. When a vulnerability is recorded, a CVSS score can be calculated. It is important to note that CVSS does not provide advice on the risk of found vulnerabilities; this limitation can be resolved using a different vulnerability framework, such as Nessus, which uses vulnerability scanners to provide risk ratings.

FIGURE 1: CVSS METRICS AND CALCULATIONS



To implement the CVSS methodology, our Canberra-based Australian cyber security testing team remotely connected to a facility in the US state of Pennsylvania, where IP cameras were stored for penetration testing. Official permission to conduct the testing was sought and received from IPVM prior to the commencement of testing. Complete details of our application of the CVSS methodology, including the IoT IP camera setup, analysis, experimental setup, software tools, information collection, and penetration testing, appear in our technical paper (Bernot et al. 2025).

B. CVSS Findings: Attacks and Exploitation of Reported Vulnerabilities

In CVSS testing, various vulnerabilities can be exploited. We tested three systems: Dahua, Hikvision, and Avigilon. We performed penetration testing using three basic cyber security attacks: denial of service (DoS), man-in-the-middle (MITM), and local area network denial (LAND).

The CVE dictionary entries for these attacks used active CVEs that were known to the suppliers but did not have fixes at the time of the attacks. First, DoS tests were run to check traffic filtering using a trusted login; all three providers fell victim to the DoS attacks, causing video streaming to stop. Second, we ran the MITM via IP spoofing that allowed the interception of network traffic using BetterCap in Kali Linux. All the IP cameras succumbed to the MITM attacks. Finally, we ran LAND attacks that tested whether the cameras could be disabled via pings redirected to the source of the camera packet sending. LAND attacks stopped the camera streams of all three providers.

We expected that the devices would be able to identify and rectify the attacks based on implemented Domain Name System (DNS) policies. However, we found these attacks to be successful even at a basic level. The second means of comparing the cyber security of these IP cameras was to compare them across a range of CVEs.

C. CVSS Findings: Fixes of Previously Reported Vulnerabilities

This section examines manufacturers' responses to reported CVEs. When evaluating CVSS findings, it is crucial to assess both the exploitation of vulnerabilities and the manufacturers' responses. Prompt action and manufacturer-initiated updates can prevent widespread exploitation of a reported CVE, but patches may not fully secure devices if the vulnerability has already been exploited. Attacks can be executed in their original forms, such as DoS, DDoS, or MITM, or by exploiting existing vulnerabilities identified in our network scans. At the time of writing, Avigilon had no known critical vulnerabilities, which reduces its attack surface. By contrast, Hikvision and Dahua have publicly known critical vulnerabilities, making them more susceptible to large-scale attacks, such as the Slowloris DoS vulnerability in Hikvision (see Figure 2, this section).

The National Vulnerability Database (NVD) is the US government repository of standards-based vulnerability management that uses CVSS protocols. The NVD records all CVEs and tracks their exploitation records. A report on a critical vulnerability—a CVE calculated at a 9–10 severity score range—should prompt rapid action by the software owner. An NVD search in December 2024 showed that Avigilon has three recorded vulnerabilities (none calculated as critical), Hikvision has 27 recorded vulnerabilities (8 critical) with one exploited vulnerability, and Dahua has 55 recorded vulnerabilities (11 critical) with two exploited vulnerabilities (NIST n.d.).

While the results do not refer exclusively to cameras, they are specific to IoT devices, many of which are likely to share the system design (NIST n.d.). A summary of the critical vulnerabilities is shown in Table II below.

TABLE II: A SUMMARY OF CRITICAL AND EXPLOITED HIKVISION AND DAHUA CVES

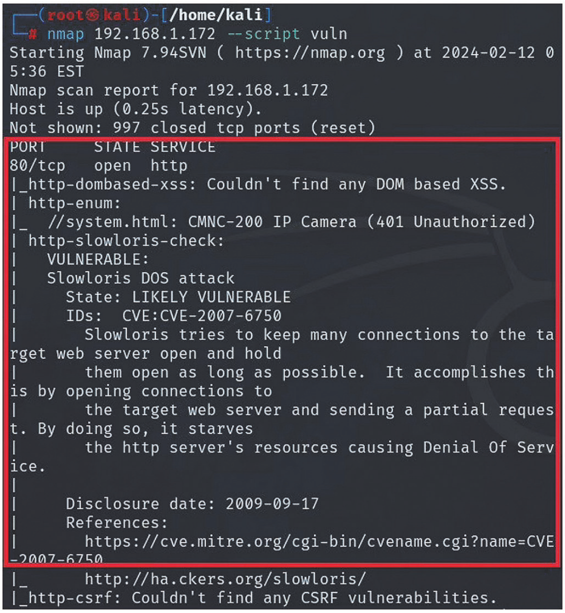
Dictionary Entry and Company	Company Associated	CVSS Severity	Description
CVE-2021-36260	Hikvision	9.8	A command injection vulnerability in the web server of some Hikvision products. Due to insufficient input validation, an attacker can exploit the vulnerability to launch a command injection attack by sending some messages with malicious commands.
CVE-2021-33045	Dahua	9.8	The identity authentication bypass vulnerability found in some Dahua products during the login process. Attackers can bypass device identity authentication by constructing malicious data packets.
CVE-2021-33044	Dahua	9.8	The identity authentication bypass vulnerability found in some Dahua products during the login process. Attackers can bypass device identity authentication by constructing malicious data packets.

Both Hikvision and Dahua reported critical vulnerabilities with a severity score of 9.8. Hikvision’s CVE-2021-36260 was reported in July 2021, and the company released an official security update on their website in September 2021. However, they did not prompt automatic firmware updates. This vulnerability has since been widely circulated and exploited and has been among those most used by Chinese state-sponsored cyber actors since 2020. In August 2022, over 80,000 Hikvision cameras were still unpatched and vulnerable due to the lack of automated firmware updates (CYFIRMA 2022). Additionally, a joint cyber security advisory from Australia, Canada, the United Kingdom, the United States, and other countries stated that hackers linked to the People’s Republic of China spied on these nations using Cisco Routers and Hikvision cameras, including 2,400 devices in Australia (Federal Bureau of Investigation, Cyber National Mission Force, and National Security Agency 2024).

The US Federal Bureau of Investigation has warned of remote exploits of Dahua and Hikvision cameras based on previous vulnerabilities (IPVM 2024a). Dahua’s critical CVEs are listed in the Known Exploited Vulnerabilities Catalog of the Cybersecurity and Infrastructure Security Agency (CISA), with details of these exploits publicly available (CISA 2024; IPVM 2024b; mcw0 2021). Our network scanning confirmed these findings. For example, as shown in Figure 2 below, the Nmap scan of the Hikvision Turret camera under investigation revealed that the Hikvision vulnerability

CVE-2007-6750, published by NVD in December 2011 and modified in November 2024, was still exploitable. CVE-2007-6750 is associated with the Slowloris DoS Attack, which targets a Web server by opening and maintaining multiple connections, thus allowing a single machine to take down the server with minimal bandwidth. Despite recent mitigation strategies like increased server scalability, rate limiting, and cloud-based reverse proxies, the Hikvision Nmap scan in Figure 2 shows it remains exploitable. A similar investigation of Dahua revealed critical authentication bypass vulnerabilities, CVE-2021-33044 and CVE-2021-33045, allowing attackers to access devices without proper credentials by sending malicious data packets. CISA noted active exploitation of these vulnerabilities in August 2024 (CISA 2024).

FIGURE 2: NMAP VULNERABILITY SCAN FOR HIKVISION TURRET CAMERA



```
┌──(root㉿kali)-[/home/kali]
└─# nmap 192.168.1.172 --script vuln
Starting Nmap 7.94SVN ( https://nmap.org ) at 2024-02-12 0
5:36 EST
Nmap scan report for 192.168.1.172
Host is up (0.25s latency).
Not shown: 997 closed tcp ports (reset)
PORT     STATE SERVICE
80/tcp   open  http
|_http-dombased-xss: Couldn't find any DOM based XSS.
| http-enum:
|_  //system.html: CMNC-200 IP Camera (401 Unauthorized)
| http-slowloris-check:
|   VULNERABLE:
|   Slowloris DOS attack
|     State: LIKELY VULNERABLE
|     IDs:  CVE:CVE-2007-6750
|       Slowloris tries to keep many connections to the ta
rget web server open and hold
|       them open as long as possible.  It accomplishes th
is by opening connections to
|       the target web server and sending a partial reques
t. By doing so, it starves
|       the http server's resources causing Denial Of Serv
ice.
|
|     Disclosure date: 2009-09-17
|     References:
|       https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE
-2007-6750
|_      http://ha.ckers.org/slowloris/
|_http-csrf: Couldn't find any CSRF vulnerabilities.
```

Avigilon has also faced several security vulnerabilities in its products over the years. For example, CVE-2021-38701 affected the administrative user interface of specific devices, allowing attackers to execute arbitrary scripts in the user's browser session. Additionally, CVE-2015-2860 allowed remote attackers to read arbitrary files on the system by sending crafted requests. Avigilon has assessed the impact of widely publicized vulnerabilities on its products, including CVE-2021-44228, CVE-2022-26809, and CVE-2022-22965. For CVE-2021-44228, Avigilon evaluated its products for exposure to the Apache Log4j2 vulnerability and provided guidance on affected and unaffected products. For CVE-2022-26809 and CVE-2022-22965, Avigilon issued notifications regarding the impact and recommended mitigation steps.

In 2024, IPVM released a report evaluating the cyber security of IP cameras based on various testing criteria (IPVM 2024c). This report highlighted key features and their availability in different IP camera models, with detailed findings available on the IPVM website. The report ranked IP camera manufacturers based on cyber security performance, revealing significant disparities. Avigilon scored above average in attack surface categories and fundamental built-in security features. Dahua also performed well in fundamental built-in security features but is vulnerable to the TLS Secure Client-Initiated Renegotiation DoS attack, lacks TLS v1.3 support, and has multiple discovery protocols enabled by default. Hikvision's Cloud/P2P IP Camera Connection showed strong results, but its Hik-Connect app had notable persistent cyber security issues.

These issues highlight the fact that effective firmware management is essential for maintaining camera security. While IPVM's testing utilized the latest firmware versions, some cameras required complex upgrade procedures, often necessitating proprietary tools and in-person updates. These complexities can delay firmware updates, leaving systems vulnerable. Consequently, identified vulnerabilities make Hikvision and Dahua cameras more susceptible to attacks.

3. THE POLICY ANALYSIS: CYBER VULNERABILITIES

The second part of our methodology consisted of policy analysis. Specifically, we considered whether the vulnerabilities identified in the first part of the research could be covered by the cyber security guidelines and regulations currently in place in Australia. The findings indicate that Australia has overlapping cyber hygiene guidelines and limited technical cyber security guidelines, which are provided by intelligence agencies to arms of the federal government only on request. The lack of mandatory technical cyber security regulation (such as the EU's CRA) leaves regulatory gaps at the level of state governments and the consumer market.

A. Overview of Australian Frameworks and Guidance for IoT IP Cameras

We reviewed the national standards and guidelines related to IoT IP camera security, which include international standards, security frameworks, mitigation strategies, information security manuals, and design principles. Three researchers from technical and regulatory academic backgrounds analyzed the aforementioned documents and corroborated the findings.

With some exceptions, these documents have one main aim: to raise the level of organizational and environmental cyber hygiene in government agencies (see Table III). This aim is directed towards hardening best practices rather than proactively overseeing supply chain regulation (e.g., establishing and monitoring minimum technical security requirements for IP camera manufacturers). Additionally, the consumer market, which holds most IP cameras in Australia, is only governed by basic cyber security requirements.

TABLE III: AN OVERVIEW OF AUSTRALIAN FRAMEWORKS/GUIDANCE FOR IOT IP CAMERAS

Framework/ Guidance/ Standard	Focus Area	Audience	Purpose
Standards Australia: General IoT and CCTV Regulations	Security and privacy standards for IoT devices and CCTV systems. A complete overview is available from Standards Australia: www.standards.org.au	Manufacturers, government agencies, and critical infrastructure operators.	Sets *baseline standards* to ensure IoT and CCTV devices are secure and compliant.
ASD's 13 Security-by-Design Principles	Secure design and development principles for ICT, including IoT devices such as CCTV cameras, drones, solar inverters and other smart devices.	Developers, system architects, and organizations managing ICT systems.	Promotes *secure-by-design* practices to embed security into systems from inception.
Essential Eight Mitigation Strategies	Key strategies to mitigate cyber threats and protect IT networks, including prevention, limiting incidents, and recovery.	Australian organizations, particularly government and businesses.	A practical guide to implement *eight core mitigation strategie*s for cyber threat resilience.
Information Security Manual (ISM) (Dec. 2024)	Cyber security framework to protect systems, data, and networks from cyber threats.	Australian government agencies, private sector, and critical infrastructure.	Provides cyber security controls to manage and mitigate evolving cyber risks.
The Protective Security Policy Framework (PSPF) 2024	A risk-based approach to managing protective security across governance, including personnel, physical, and information security.	Australian government entities.	Establishes *security standards and guidelines* to safeguard people, information, and assets.
ASIO-T4 Protective Security Advice	Physical and protective security advice to safeguard people, assets, and facilities.	State and territory government agencies, law enforcement, critical infrastructure, and national security entities.	Offers guidance on *physical security measures* to protect against national security threats.

Standards Recommended for Implementation			
Standard	**Focus Area**	**Audience**	**Purpose**
ETSI EN 303 645 (Recommendation by the Office of Impact Analysis and Engineers Australia)	Cyber security standard for consumer IoT devices, such as IP cameras, wearable health devices, and home automation and alarm systems.	IoT manufacturers, policymakers, and organizations adopting IoT.	Outlines *essential security requirements* for consumer IoT products to reduce vulnerabilities.
IEC 62443 (Recommendation by Engineers Australia)	Cyber security standard for industrial automation and control systems (IACS).	Critical infrastructure operators and industrial control system (ICS) providers.	Provides a *framework for securing industrial systems* against cyber threats.

In addition to the regulatory documents, the Australian government has also issued advice in response to specific attacks. In May 2024, Australia and its Five Eyes intelligence partners publicly attributed cyber attacks on US infrastructure to China, naming “Volt Typhoon” as the state-sponsored cyber actor (Australian Cyber Security Centre 2022). The Australian Signals Directorate published a fact sheet for security leaders to establish strong vendor risk management processes. This includes using non-binding guidance documents, ensuring vendors have a patching plan, and limiting product usage that violates the principle of least privilege (Australian Cyber Security Centre 2022).

B. Overlapping Cyber Hygiene Guidelines and Lacking Technical Regulations

Australia has multiple security frameworks, guidelines, and regulations that cover IoT IP camera security, mainly focusing on organizational cyber hygiene and environmental security. Because the country’s technology market is small, it has always been difficult to develop new national technology guidelines. The article notes a lack of regulations addressing the high and critical vulnerabilities found in Hikvision and Dahua cameras, suggesting that IoT IP cameras are still under-regulated from a technical perspective. Therefore, we argue that Australia’s IoT devices are both over- and under-regulated. Specifically, organizational cyber hygiene requirements, as outlined in the Protective Security Policy Framework and the Information Security Manual, are mandatory for government agencies and some critical infrastructure businesses. However, only basic technical security requirements are currently implemented for IP camera manufacturers.

We also find that voluntary compliance is a significant issue with Australia’s regulatory documents. Among the national regulatory documents analyzed, only the Standards

Australia baseline standards for IoT and video surveillance devices are mandatory. Outside of these standards, only Australian federal government agencies must comply with mandatory cyber security guidelines. Namely, the Information Security Manual, the Protective Security Policy Framework, and the Essential Eight mitigation strategies are mandatory and audited. The voluntary nature of the frameworks/guidance documents means they are rarely used in practice and weakens the security of IoT devices. While industry self-regulation can help mitigate emerging harms and establish best practices, it is unlikely to support the enforcement of these practices (Gunningham and Rees 1997). As observed in the previous section, this regulatory gap particularly harms the consumer market, where cost-competitive devices perform better, forcing manufacturers into price-based competition that discourages a focus on cyber security.

In December 2023, the Office of Impact Analysis (OIA) responded to the Australian government's plan to co-design mandatory cyber security standards as part of the 2023–2030 Australian Cyber Security Strategy. The OIA ensures that policy decisions are evidence-based by assessing their economic and social impacts, providing guidance, and promoting transparency. In their proposal, the OIA acknowledged the absence of mandatory security standards and addressed the associated harms (OIA 2023):

> At present, smart device manufacturers are not required to comply with security standards, which can lead to an increased risk of vulnerability that may be exploited, exposing consumers to cyber risks. Due to persistent and preventable cyber security incidents, Australian households and businesses are bearing financial costs and negative societal impacts. Estimates of these costs are as high as $29 billion per year. Consumers often cannot distinguish between a secure and insecure device due to a lack of clear and accessible information. This limits commercial incentives for manufacturers to prioritize security, leading to consumers unknowingly adopting cyber security risk.

The OIA supported two initiatives that align with the goals of the Australian 2023–2030 Cyber Security Strategy—the adoption of international standards and a consumer labeling scheme. The recommended standard is the European Telecommunications Standards Institute (ETSI) Cyber Security for Consumer Internet of Things: Baseline Requirements (ETSI EN 303 645). The three baseline principles of the standard are no default universal passwords, a vulnerability disclosure policy, and software updates. In addition to the proposed adoption of international standards, the IoT IP camera labeling scheme has been introduced as a solution for consumers. Early consumer research shows that it could be effective. In a 2022 representative survey, almost 20%

of consumers agreed they were more likely to buy a device with a graded security label than a device with no label (Tonkin 2022).

Our research findings show that without mandatory basic security requirements, vulnerabilities will continue to be present in IoT IP cameras. Namely, our technical findings showed that even basic cyber attacks were successful. Additionally, some vulnerabilities were not fixed by manufacturers even after they had been reported and researched, affecting numerous devices. This implies that without a comprehensive cyber security uplift, including mandatory security guidelines, the Australian government, businesses, and consumers will respond to IoT vulnerabilities only reactively. A delayed response would necessitate a continued band-aid approach to cyber security as intelligence agencies identify national security risk priorities.

C. Effects of Under-Regulation on National and Private Security

This article highlights manufacturers' cyber security practices as the main issue and argues that a national cyber security uplift for IoT technologies is required. The under-regulation of IoT IP cameras allows insecure technologies to persist in the Australian market. Even if security risks to federal buildings are mitigated by replacing Hikvision and Dahua cameras with the more secure Avigilon alternatives, many Hikvision and Dahua cameras remain deployed nationwide. Therefore, we argue that Australia's federal-building camera-replacement efforts offer little benefit to everyday Australians and serve only as a temporary solution to long-term cyber security risks.

Security uplift options should ideally extend to the whole of government as well as consumers. In the current regulatory context, consumers have the most to lose and little recourse for reporting issues or complaints. Broadcast hacks linked to Russian IP addresses streamed camera footage from Australian businesses and homes in 2020 (Roberts 2020) and in October 2024 (AUCyber 2024). In 2021, a group of hackers claimed to have accessed Verkada cameras with facial recognition capabilities used by more than 100 Australian organizations, including childcare centers, schools, and aged care (Purtill 2021). To mitigate these risks, a cyber security uplift is required for IoT surveillance devices, especially those available on the consumer market, as video streams from IP cameras can be and have been used to facilitate crime and social harms.

The Cyber Resilience Act (CRA) passed by the EU in December 2024 provides an example of what such regulation could look like. The CRA is one of the world's first regulations that explicitly target IoT devices with the goal of reducing cyber incidents by placing more responsibility on IoT device manufacturers and distributors. When the CRA is fully implemented, manufacturers' obligations will include addressing cyber security risks in all phases (planning, design, development, production, delivery, and maintenance), documenting these risks, reporting actively exploited vulnerabilities and incidents, ensuring effective handling of vulnerabilities throughout the support period, providing clear instructions for product use, and making security updates available for the product's expected duration of use. The CRA motivates compliance by imposing large financial penalties in case of breaches to the act. Another example is India's Electronics and Information Technology Goods Order, which is set to come into effect in April 2025. It will work on a smaller scale by mandating cyber security checks for imported IP cameras before they are sold in India.

The Australian government could similarly prioritize a cyber security uplift to ensure that all IoT devices on the market—not just China-made IP cameras—meet high security standards, thereby protecting national security as well as the security of everyday Australians. This could include mandatory security certifications for manufacturers and regular security audits for the lifetime of IoT products supplied to the market. As a small technology market, Australia can also consider adopting international standards and making them mandatory for manufacturers to comply with. The simultaneous adoption of a labeling scheme can support the communication of IP camera security findings to consumers. Consumer research shows they can effectively prompt individuals to choose more secure products without requiring them to obtain specialist knowledge (Tonkin 2022).

4. CONCLUSION

Following the dual methodology, our findings offer two conclusions. First, we find that the Avigilon cameras that replaced the more than 1,000 Hikvision and Dahua cameras on the federal buildings in Canberra carry fewer cyber security risks. This conclusion is further supported by the poor records of Huawei and Dahua in addressing reported technical vulnerabilities and exposures. Second, Australia is underprepared to counter the vulnerabilities and exposures found, as the current regulations cannot mitigate them. The overlapping security frameworks, guidelines, and regulations do address organizational cyber hygiene and environmental security but focus little on technical standards.

The Australian case study suggests that guidelines on organizational and environmental security are not sufficient and that politicizing imported technologies only serves as a temporary band-aid solution in enhancing cyber security. While proactive practices by manufacturers to maintain their products' cyber security are key (e.g., proactive red-teaming), regulations can support cybersecure manufacturing practices, as demonstrated by the EU's Cyber Resiliency Act 2024 for fostering IoT resiliency. Establishing security standards is essential for ensuring the safety of IoT IP cameras in government, business premises, and individual homes. Effective regulatory approaches should emphasize the role of manufacturers and distributors in launching secure products to both consumer and professional security markets.

REFERENCES

AIVD. 2022. “Threat Assessment State-sponsored Actors 2 (TASA).” November 2022. https://english.aivd.nl/publications/publications/2023/02/13/threat-assessment-state-sponsored-actors-2-tasa.

Alam, Tanweer. 2018. “A Reliable Communication Framework and Its Use in Internet of Things (IoT).” *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 5 (10).

Antonakakis, Manos, Tim April, Michael Bailey, Matthew Bernhard, Elie Bursztein, Jaime Cochran, Zakir Durumeric, et al. 2017. “Understanding the Mirai Botnet.” In *Proceedings of the 26th USENIX Security Symposium (USENIX Security 2017)*, 1093–1110. https://www.usenix.org/system/files/conference/usenixsecurity17/sec17-antonakakis.pdf.

AUCyber. 2024. “AUCyber Warning: Aussie Security Cameras at Risk | AUCyber.” *AUCloud (blog)*, October 9, 2024. https://aucyber.com.au/news/aus-security-cameras-at-risk/.

Australian Cyber Security Centre. 2022. “PRC State-Sponsored Cyber Activity: Actions for Critical Infrastructure Leaders.” December 15, 2022. https://www.cyber.gov.au/about-us/view-all-content/alerts-and-advisories/prc-state-sponsored-cyber-activity_actions-for-critical-infrastructure-leaders.

Bernot, Ausma, and Marcus Smith. 2023. “Understanding the Risks of China-Made CCTV Surveillance Cameras in Australia.” *Australian Journal of International Affairs* 77 (4). https://doi.org/10.1080/10357718.2023.2248915.

Bernot, Ausma, Muhammad Arif Khan, Khurram Shahzad, Mert Karakaya, and Conor Healy. 2025. “Cyber Vulnerabilities and Technical Regulation of China-made IoT Surveillance Cameras in Australia.” SSRN. Posted February 24, 2025. https://papers.ssrn.com/abstract=5079545.

Blythe, John M., and Shane D. Johnson. 2021. “A Systematic Review of Crime Facilitated by the Consumer Internet of Things.” *Security Journal* 34 (1). https://doi.org/10.1057/s41284-019-00211-8.

Brown, Andi, Diarmaid Harkin, and Leonie Maria Tanczer. 2024. “Safeguarding the ‘Internet of Things’ for Victim-Survivors of Domestic and Family Violence: Anticipating Exploitative Use and Encouraging Safety-by-Design.” *Violence Against Women* 31 (5). https://doi.org/10.1177/10778012231222486.

Chen, Yu-fu. 2022. “Report Details Why Chinese Products Banned at Agencies.” *Taipei Times*, October 25, 2022. https://www.taipeitimes.com/News/taiwan/archives/2022/10/25/2003787690.

CUJOAI. 2021. “15 Most Popular IoT Products and Devices in 2021.” July 27, 2021. https://cujo.com/blog/15-most-popular-iot-products-and-devices-in-2021.

Cybersecurity and Infrastructure Security Agency (CISA). 2024. “CISA Adds Four Known Exploited Vulnerabilities to Catalog.” August 21, 2024. https://www.cisa.gov/news-events/alerts/2024/08/21/cisa-adds-four-known-exploited-vulnerabilities-catalog.

CYFIRMA. 2022. “Thousands of Hikvision Cameras Are Still Vulnerable and Can Be Potentially Exploited.” August 21, 2022. https://www.cyfirma.com/research/thousands-of-hikvision-cameras-are-still-vulnerable-and-can-be-potentially-exploited/.

ENISA. 2024. “Cyber Europe 2024—After Action Report.” December 10, 2024. https://www.enisa.europa.eu/publications/cyber-europe-2024-after-action-report.

Federal Bureau of Investigation, Cyber National Mission Force, and National Security Agency. 2024. “People’s Republic of China-Linked Actors Compromise Routers and IoT Devices for Botnet Operations.” September 18, 2024. https://media.defense.gov/2024/Sep/18/2003547016/-1/-1/0/CSA-PRC-LINKED-ACTORS-BOTNET.PDF.

Federal Communications Commission. 2022. “FCC Bans Authorizations for Devices That Pose National Security Threat.” November 25, 2022. https://www.fcc.gov/document/fcc-bans-authorizations-devices-pose-national-security-threat.

Gunningham, Neil, and Joseph Rees. 1997. “Industry Self-Regulation: An Institutional Perspective.” *Law and Policy* 19 (4). https://doi.org/10.1111/1467-9930.t01-1-00033.

Guo, Y., C. Du, Z. Mustafaoglu, A. Sengur, H. Garg, K. Polat, and D. Koundal. 2025. “Using Hybrid Transformer and Convolutional Neural Network for Malware Detection in Internet of Things.” *International Journal of Pattern Recognition and Artificial Intelligence* 39 (2): 2550002. https://doi.org/10.1142/S0218001425500028.

IoT Security Foundation. 2021. “Consumer IoT Sector—Basic Cybersecurity Hygiene Practice Still Not Happening.” November 4, 2021. https://iotsecurityfoundation.org/consumer-iot-sector-basic-hygiene-practice-still-not-happening/.

IPVM. 2024a. “Detailed Report on FBI Warnings Regarding Hikvision and Dahua.” December 24, 2024. https://s.ipvm.com/uploads/embedded_file/e398ff559b32195e587f07b0a7f8c5e6057e50f003b7ee320d0925adfa937d0d/3e60acdf-a18e-48d3-b872-0723cbe23580.pdf.

IPVM. 2024b. “CISA Warns About Dahua Security Flaws.” September 18, 2024. https://ipvm.com/reports/cisa-dahua-2024.

IPVM. 2024c. “Cybersecurity Rankings Criteria for IP Cameras and NVRs Explained.” February 9, 2024. https://ipvm.com/reports/cam-cyb-criteria.

Lauritzen, Daniel Bue. 2024. “The Danish Road Directorate Waves Goodbye to Chinese Camera Equipment to Strengthen Cybersecurity.” *Altinget*, August 8, 2024. https://www.altinget.dk/artikel/vejdirektoratet-rydder-op-i-kameraudstyr-for-at-styrke-cybersikkerheden.

Mcw0. 2021. “Dahua Backdoor Proof of Concept.” September 6, 2021. https://github.com/mcw0/PoC/blob/master/dahua-backdoor-PoC.py.

Mell, Peter, Karen Scarfone, and Sasha Romanosky. 2007. The Common Vulnerability Scoring System (CVSS) and Its Applicability to Federal Agency Systems. National Institute of Standards and Technology.

Migliano, Simon, and Samuel Woodhams. 2021. “Hikvision and Dahua Surveillance Cameras: Global Locations.” Top10VPN, last updated November 16, 2021. https://www.top10vpn.com/research/hikvision-dahua-surveillance-cameras-global-locations/.

MITRE. 2021. “CVE-2021-36260.” https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2021-36260.

MITRE. 2025. “CVE—Common Vulnerabilities and Exposures.” https://cve.mitre.org/.

National Security Agency (NSA), Cybersecurity and Infrastructure Security Agency (CISA), and Federal Bureau of Investigation (FBI). 2022. “Top CVEs Exploited by PRC Cyber Actors.” October 6, 2022. https://media.defense.gov/2022/oct/06/2003092365/-1/-1/0/joint_csa_top_cves_exploited_by_prc_cyber_actors_.pdf.

NIST. n.d. “National Vulnerability Database (NVD).” Accessed December 16, 2024. https://nvd.nist.gov/.

NKSC. 2020. “Cybersecurity Assessment of Hikvision and Dahua Cameras.” May 27, 2020. https://www.nksc.lt/doc/biuleteniai/2020-05-27%20Hikvision%20ir%20Dahua%20kameru%20kibernetinio%20saugumo%20vertinimas.pdf.

NL Times. 2024. “Amsterdam Replacing Chinese CCTV and Traffic Cameras over Spying and Human Rights Concerns.” June 10, 2024. https://nltimes.nl/2024/06/10/amsterdam-replacing-chinese-cctv-traffic-cameras-spying-human-rights-concerns.

Office of Impact Analysis (OIA). 2023. “Mandatory Security Standards and Industry-Led Voluntary Cyber Security Labelling Scheme for Consumer-Grade Smart Devices.” December 11, 2023. https://oia.pmc.gov.au/published-impact-analyses-and-reports/mandatory-security-standards-and-industry-led-voluntary-cyber.

Open Web Application Security Project (OWASP). n.d. “OWASP Internet of Things.” Accessed December 18, 2024. https://owasp.org/www-project-internet-of-things/.

Palo Alto Networks. 2021. “Are the Security Cameras in Your Organization Safe from Cyber Attacks?” March 15, 2021. https://www.paloaltonetworks.com/blog/network-security/are-your-security-cameras-safe-from-cyberattacks/.

Purtill, James. 2021. “Hackers Say They’ve Gained Access to Surveillance Cameras in Australian Childcare Centres, Schools and Aged Care.” *ABC News*, March 11, 2021. https://www.abc.net.au/news/science/2021-03-11/verkada-hackers-gained-access-to-australian-surveillance-cameras/13237820.

Roberts, George. 2020. “Australian Security Cameras Hacked, Streamed on a Russian-Based Website.” *ABC News*, June 24, 2020. https://www.abc.net.au/news/2020-06-24/security-cameras-hacked-streamed-on-russian-website/12380606.

Scanlan, John, and Conor Healy. 2024. “Honeywell Hides OEMing PRC China Sunell.” *IPVM*, August 19, 2024. https://ipvm.com/reports/honeywell-oems-sunell.

Sinha, Satyajit. 2024. “State of IoT 2024: Number of Connected IoT Devices Growing 13% to 18.8 Billion Globally.” September 3, 2024. https://iot-analytics.com/number-connected-iot-devices/.

Siwakoti, Yuba Raj, Manish Bhurtel, Danda B. Rawat, Adam Oest, and R. C. Johnson. 2023. “Advances in IoT Security: Vulnerabilities, Enabled Criminal Services, Attacks, and Countermeasures.” *IEEE Internet of Things Journal* 10 (13). https://doi.org/10.1109/JIOT.2023.3252594.

Tonkin, Casey. 2022. “Smart Devices Should Have a Security Health Label.” *Information Age*, May 3, 2022. https://ia.acs.org.au/article/2022/smart-devices-should-have-a-security-health-label.html.

CyCon 2025: The Next Step
17th International Conference on Cyber Conflict
C. Kwan, N. Gratzer, K. Podiņš, M. Tolppa (Eds.)
2025 © NATO CCDCOE Publications, Tallinn

Permission to make digital or hard copies of this publication for internal use within NATO and for personal or educational use when for non-profit or non-commercial purposes is granted provided that copies bear this notice and a full citation on the first page. Any other reproduction or transmission requires prior written permission by NATO CCDCOE.

Vulnerability Patch Verification for Military Software Systems Through AI-Driven Code-Level Rule Generation

Siam Shibly Antar
Research Assistant
School of Computing
Queen's University
Kingston, ON, Canada
siamshibly.antar@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: Patch verification is critical in military systems to ensure that known vulnerabilities are effectively addressed, preventing them from being exploited. Without proper verification, unpatched software could allow adversaries to exploit vulnerabilities, leading to unauthorized access, compromised operations, or even mission failure. In high-stakes environments such as military operations, patch verification is essential for maintaining the security, integrity, and readiness of both software and firmware, particularly in systems that manage sensitive information or control mission-critical equipment.

Traditional methods that rely on version strings to verify vulnerability patching are often insufficient. For example, the Heartbleed vulnerability (CVE-2014-0160) affected OpenSSL versions 1.0.1 through 1.0.1f. A system running OpenSSL 1.0.1f might still be flagged as vulnerable, even if a custom patch was applied, in the event that the version string was not updated by the software maintainer fixing the vulnerability. This will lead to false positives in the vulnerability detection process. Conversely, a system may appear secure based on the version string, but if the patch was not correctly implemented, the vulnerability will remain, resulting in false negatives. To address these limitations, this paper presents a new scalable, artificial intelligence-based code-level verification system. By leveraging large language models to generate rules that analyze the actual executable code, this approach verifies whether vulnerabilities have been properly fixed, regardless of version metadata. Additionally, it can pinpoint the exact location of exploitable code as a more accurate and reliable method for detecting and confirming patches. Our experiment, involving 1,466 vulnerable software records with over 4,000 instances, demonstrates that the rule generation system is both accurate and robust.

Keywords: *patch verification, vulnerability detection, firmware analysis, artificial intelligence, large language models*

1. INTRODUCTION

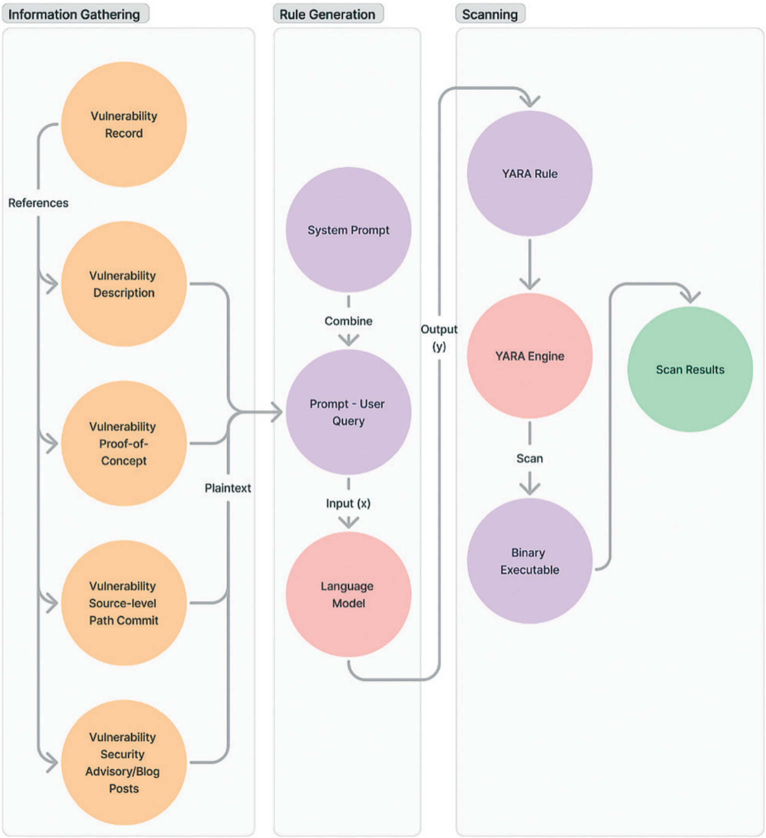
Vulnerability patch verification is a critical process in maintaining the security and reliability of software systems, particularly in high-stakes environments such as military operations. It ensures that known vulnerabilities, such as newly disclosed common vulnerabilities and exposures (CVEs) or any in-house new vulnerability records, have been effectively addressed or evaluated to mitigate potential risks. For example, when a new vulnerability is publicly announced through a CVE or found by internal threat teams, organizations must rapidly assess the associated risks and confirm that their systems are not vulnerable. This process is important for preventing unauthorized access, data breaches, and operational disruptions that could compromise mission-critical systems. An example of this urgency was seen with the Shellshock vulnerability (CVE-2014-6271), where exploits targeting this Bash flaw appeared within hours of its disclosure, affecting millions of devices reliant on Bash for system-level operations [1]. Even after almost 10 years, it remains one of the most exploited vulnerabilities, despite a patch being available [2]. This underscores the importance of swift response and thorough verification of newly discovered vulnerabilities.

Traditional methods of vulnerability verification often rely on information provided by software vendors. This approach is fraught with challenges. The complexity of modern software supply chains, combined with the increasing prevalence of supply chain attacks—such as the SolarWinds attack [3], which exploited software updates to introduce malicious code into thousands of organizations, or the attack on Kaseya

VSA [4], where a compromised IT management tool led to widespread ransomware infections—undermine the reliability of vendor-provided data. A one-day or multi-day delay in vendors validating and addressing newly disclosed CVEs can leave systems exposed. Additionally, open-source software (OSS) further complicates this landscape [5], as its decentralized nature can lead to inconsistencies in patch deployment and versioning. An example would be the Log4j vulnerability (CVE-2021-44228), which impacted millions of devices globally [6]. This flaw in the widely used Log4j library, part of the Apache Logging Services, exposed systems to remote code execution attacks, highlighting how a single OSS vulnerability can have widespread consequences. Vendors may customize OSS libraries or fail to update version strings, making it difficult to determine whether a vulnerability has been patched. Vendor information cannot always be trusted.

To address these shortcomings, a zero-trust approach has gained traction, involving three key steps: generating a software bill of materials (SBOM) [1], monitoring new CVEs that match the SBOM catalog, and analyzing systems for unpatched vulnerabilities. An SBOM provides a detailed inventory of software components, including their origins, versions, and dependencies, enabling organizations to map CVEs to their systems. For example, consider a system running OpenSSL with the version string "OpenSSL 1.0.1f." This version string can be linked to product details, allowing tools to identify known vulnerabilities such as the Heartbleed vulnerability (CVE-2014-0160), which affects OpenSSL versions 1.0.1 through 1.0.1f [2]. FACT [7], EMBA [8], CVE Binary Tool [9], and ERS0 [10] follow such an approach. However, reliance on version string matching introduces significant risks. Vendors may adopt OSS or software from sub-vendors with altered version string patterns, complicating the identification process. Additionally, vendors may implement in-house patches for CVEs without updating the version string, especially when other parts of the code remain unchanged. A patched version of OpenSSL, for instance, might still appear vulnerable if the version string remains unmodified. Conversely, systems might seem secure based on metadata, while still harboring unpatched vulnerabilities due to incomplete fixes or custom versions. These limitations highlight the need for a more robust and precise method.

FIGURE 1: THE PROCESS BEGINS WITH INFORMATION GATHERING, WHICH SERVES AS THE INPUT FOR A LANGUAGE MODEL. THE GENERATED PROMPTS LEAD TO THE CREATION OF YARA RULES, WHICH ARE SUBSEQUENTLY USED FOR SCANNING BINARY FILES, CULMINATING IN THE GENERATION OF SCAN RESULTS



A novel approach bypasses these metadata-based limitations by directly verifying the presence of unpatched CVEs in software binaries (see Figure 1). Using publicly available CVE patch information, such as source code commit logs, pattern-matching rules can be generated to identify instances of unpatched vulnerabilities in executable code. YARA rules [11], a flexible and performance-optimized pattern-matching framework, have been selected for this purpose. Commonly used in malware detection and triage, YARA rules enable efficient scanning of binaries, making them well-suited for large-scale vulnerability analysis.

While YARA rules are traditionally crafted manually, this process is time-consuming and does not scale to the volume of newly released CVEs. To address this, we propose a novel approach leveraging language models to automate the generation of YARA rules for unpatched CVEs. By taking CVE information, including proof-of-concept (PoC) exploits and patch commit logs as the input, the system generates YARA rules to detect the corresponding vulnerabilities in binary executables. This automated method not only accelerates the process but also enhances explainability, as the generated rules clearly delineate where vulnerabilities exist and why they remain unpatched. Preliminary investigations reveal that existing language models struggle to produce high-quality YARA rules. To overcome this limitation, we introduce a two-phase training methodology designed to improve the quality of the generated rules. The contributions of this paper are as follows:

- We propose a fast and reliable method for vulnerability patch verification and risk assessment, adopting a zero-trust approach that does not depend on vendor-provided information.
- We present a two-phase training framework for language models to generate high-quality vulnerability detection rules conforming to YARA specifications.
- We benchmark various language models for their effectiveness in generating vulnerability-matching rules, demonstrating the efficacy of our proposed approach.

This paper is structured as follows: Section 2 reviews related works. Section 3 formally defines the research problem. Section 4 outlines our methodology for model training. Section 5 details the experimental results. Finally, Section 6 provides the conclusion.

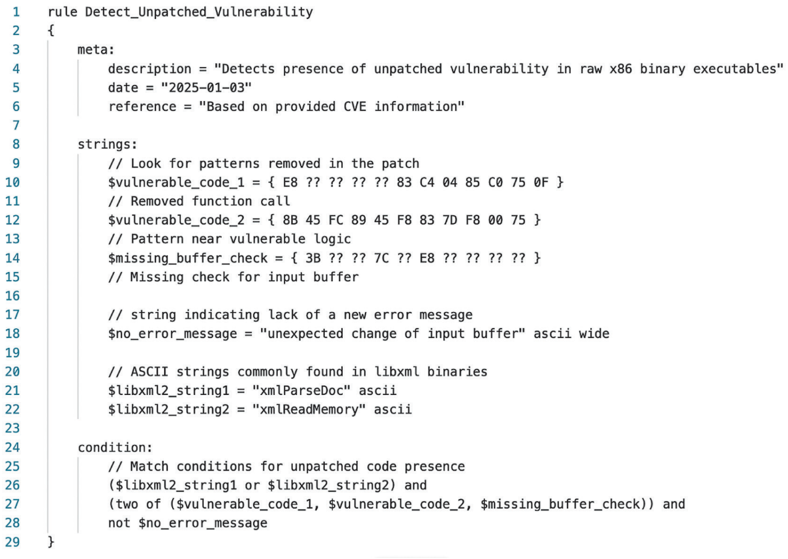
2. RELATED WORKS

Vulnerability detection involves identifying software flaws that can be exploited by attackers. It can be broadly categorized into static and dynamic approaches. Static vulnerability detection analyzes the source code, binaries, or intermediate representations without executing the program. Model-based approaches, such as taint analysis [12], track the flow of potentially malicious inputs through the program to identify insecure patterns. Data-driven methods leverage deep learning models trained on large datasets of vulnerable and non-vulnerable code to predict flaws [13]. While static methods provide comprehensive coverage, they may produce false positives due to the lack of runtime context.

Dynamic vulnerability detection, on the other hand, analyzes the software during execution to identify vulnerabilities that arise only under specific runtime conditions. Widely used techniques include fuzz testing [14], which provides random or malformed inputs to the program, and symbolic execution [15], which systematically explores execution paths. While these methods are effective in finding runtime-specific vulnerabilities, they can be resource-intensive and may miss issues that are not triggered during testing.

Vulnerability scanning, the paradigm under which this work falls, focuses on identifying known vulnerabilities within software systems. This approach often uses an SBOM to map vulnerabilities to specific components within a system [7], [8], [9], [10]. Another common method involves assembly code clone detection, which identifies code similarities to known vulnerable software [16], [17]. While assembly code clone approaches provide fuzzy matching results, typically in the form of a matching score between 0 and 1, they face challenges such as determining appropriate thresholds and requiring manual verification to finally confirm vulnerabilities. Despite these challenges, code clone techniques have advantages in identifying vulnerabilities at the binary level, but require disassembly, which can increase complexity.

FIGURE 2: EXAMPLE YARA RULE FOR DETECTING UNPATCHED CVE-2017-9049



```
rule Detect_Unpatched_Vulnerability
{
    meta:
        description = "Detects presence of unpatched vulnerability in raw x86 binary executables"
        date = "2025-01-03"
        reference = "Based on provided CVE information"

    strings:
        // Look for patterns removed in the patch
        $vulnerable_code_1 = { E8 ?? ?? ?? ?? 83 C4 04 85 C0 75 0F }
        // Removed function call
        $vulnerable_code_2 = { 8B 45 FC 89 45 F8 83 7D F8 00 75 }
        // Pattern near vulnerable logic
        $missing_buffer_check = { 3B ?? ?? 7C ?? E8 ?? ?? ?? ?? }
        // Missing check for input buffer

        // string indicating lack of a new error message
        $no_error_message = "unexpected change of input buffer" ascii wide

        // ASCII strings commonly found in libxml binaries
        $libxml2_string1 = "xmlParseDoc" ascii
        $libxml2_string2 = "xmlReadMemory" ascii

    condition:
        // Match conditions for unpatched code presence
        ($libxml2_string1 or $libxml2_string2) and
        (two of ($vulnerable_code_1, $vulnerable_code_2, $missing_buffer_check)) and
        not $no_error_message
}
```

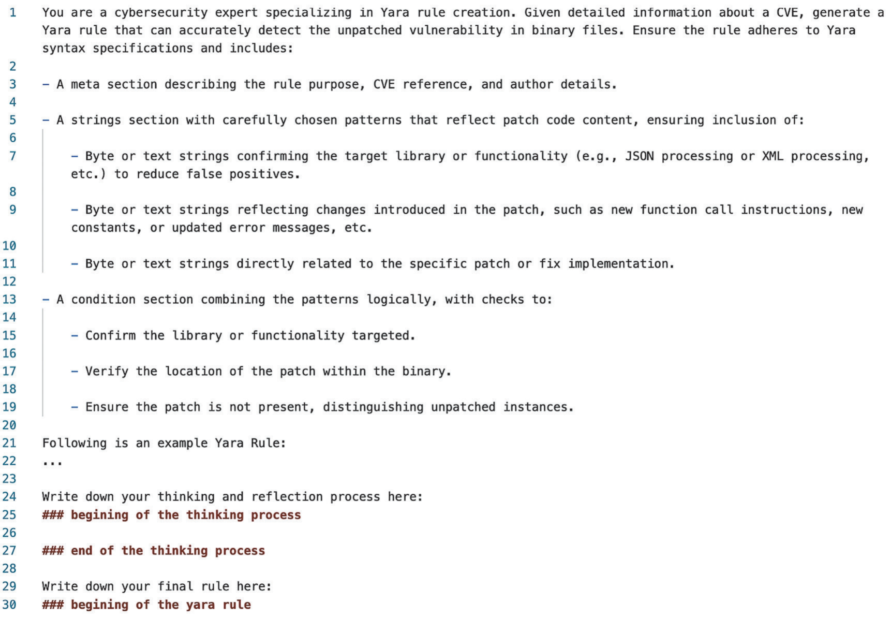
Our work diverges by emphasizing fast triage through explainable matching patterns. Unlike binary code clone approaches, which prioritize detailed matching at the cost of performance, our method focuses on generating transparent and actionable vulnerability rules. This approach balances precision and efficiency, providing a scalable method for rapid vulnerability scanning and verification as a standalone solution, or a complement to existing clone search-based methods. We are among the first to adopt this strategy, combining explainability and speed to address the challenges of vulnerability verification in a novel and effective way.

3. PROBLEM DEFINITION

The problem involves transforming vulnerability record information into actionable detection rules for binary files. The input consists of a released CVE's details, including its description and all related data available under the references section formatted as plain text and denoted as *x*. For example, on the National Institute of Standards and Technology national vulnerability database, there is a "References to Advisories, Solutions, and Tools" section for each CVE record.

This information is collected using automated crawlers that retrieve relevant details such as threat advisories, descriptions, source code commits of patches, PoC exploits, and blog posts analyzing the vulnerability. Leveraging this diverse data source ensures a comprehensive understanding of the vulnerability and its exploitation patterns for rule generation.

FIGURE 3: SYSTEM PROMPT DESIGN



```
You are a cybersecurity expert specializing in Yara rule creation. Given detailed information about a CVE, generate a
Yara rule that can accurately detect the unpatched vulnerability in binary files. Ensure the rule adheres to Yara
syntax specifications and includes:

- A meta section describing the rule purpose, CVE reference, and author details.

- A strings section with carefully chosen patterns that reflect patch code content, ensuring inclusion of:

    - Byte or text strings confirming the target library or functionality (e.g., JSON processing or XML processing,
    etc.) to reduce false positives.

    - Byte or text strings reflecting changes introduced in the patch, such as new function call instructions, new
    constants, or updated error messages, etc.

    - Byte or text strings directly related to the specific patch or fix implementation.

- A condition section combining the patterns logically, with checks to:

    - Confirm the library or functionality targeted.

    - Verify the location of the patch within the binary.

    - Ensure the patch is not present, distinguishing unpatched instances.

Following is an example Yara Rule:
...

Write down your thinking and reflection process here:
### begining of the thinking process

### end of the thinking process

Write down your final rule here:
### begining of the yara rule
```

In this paper, we focus on public records to build the required input dataset. These records include advisories from official CVE databases, Git repositories documenting patch implementations, security researchers' PoC codes (optional), and technical blogs discussing the vulnerability's scope and impact. While our approach is based on public data, the same methodology can be applied to in-house vulnerability records, where organizations can gather similar information internally through proprietary systems and sources.

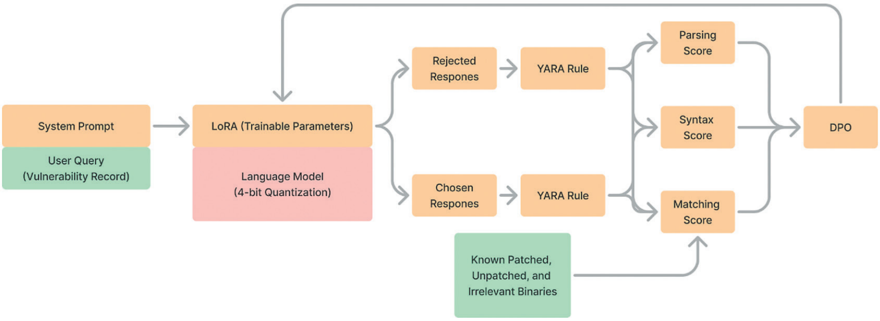
The goal is to generate a YARA rule, denoted as y, capable of identifying unpatched instances of the vulnerability in binary executables (see Figure 2). YARA rules provide explainable and precise matching patterns that facilitate rapid detection and verification of vulnerabilities across diverse systems. By automating this process, we aim to enhance scalability, while maintaining high levels of accuracy and interpretability for vulnerability detection.

4. METHODOLOGY

A. Prompt Engineering for YARA Rule Generation

The initial step starts with designing effective prompts to guide the language model in generating YARA rules (see Figure 4). Prompts typically consist of two parts: the system prompt and the user query [18]. The system prompt provides a detailed set of instructions and context for the model, such as "Generate a YARA rule for detecting a vulnerability based on the provided CVE details. Ensure the rule adheres to YARA specifications and includes meaningful identifiers and conditions." This part sets the task's scope and quality expectations. The user query, by contrast, supplies the specific input data for the task. For example, a query might state: "Based on CVE-2021-44228, generate a YARA rule. The CVE details are as follows: ### start of CVE details .... ### end of CVE details."

FIGURE 4: THE OVERALL TRAINING WORKFLOW AND REWARD SCORE CALCULATION


System Prompt
User Query (Vulnerability Record)
LoRA (Trainable Parameters)
Language Model (4-bit Quantization)
Rejected Respones
Chosen Respones
YARA Rule
YARA Rule
Known Patched, Unpatched, and Irrelevant Binaries
Parsing Score
Syntax Score
Matching Score
DPO

We first draft a base system prompt that incorporates key elements such as YARA rule structure, syntax requirements, and general considerations about rule quality. This base prompt is then iteratively refined using outputs from a separate language model. Manual feedback is employed to evaluate the generated rules for alignment with predefined standards, such as syntactic validity and contextual accuracy. This iterative refinement involves adjusting the phrasing—for instance, inclusion—and input-output formats of the prompts to optimize the model's ability to produce high-quality and consistent YARA rules. Figure 3 shows our example system prompt. Contextual information about the vulnerabilities will be used as the user query prompt.

B. Language Model Initial Setup

To enhance the efficiency and scalability of rule generation, we employ low-rank adaptation (LoRA) and 4-bit quantization (see Figure 4). These optimization methods enable the effective adaptation of pre-trained language models to the specialized tasks, in our case YARA rule generation, while minimizing computational and resource overhead. Especially for model fine-tuning, the reduced overhead enables us to train the model in faster iterations.

LoRA is a fine-tuning method that optimizes pre-trained models by injecting additional learnable parameters into low-rank matrices within specific layers of the model [19]. This approach focuses on training only the newly introduced parameters, while leaving the pre-trained weights untouched. By reducing the number of trainable parameters, LoRA significantly decreases memory and computational requirements compared to traditional fine-tuning. This makes LoRA particularly useful for tasks requiring domain-specific adaptation, such as cybersecurity applications, where the model can efficiently specialize in YARA rule generation without losing its general-purpose capabilities.

In 4-bit quantization, a model is compressed by representing its weights with 4 bits instead of the typical 16 or 32 bits, reducing the model size drastically [20]. This compression allows for faster inference times and enables deployment on hardware with limited computational power, such as edge devices or low-resource servers. Despite the reduction in precision, modern quantization techniques use algorithms to maintain the model's accuracy, ensuring that it performs well even under these constraints. For YARA rule generation, 4-bit quantization ensures that the model is efficient enough for real-time and large-scale applications in varying application scenarios.

C. Iterative Sampling for YARA Rule Syntax Correction

The language model may fail to generate syntactically correct YARA rules due to issues such as:

- Including extraneous explanation text or code snippets outside the designated response area, leading to extraction errors.
- Producing YARA rules that are not syntactically valid.

To address these challenges, we consider two methods for training the existing language model: direct preference optimization (DPO) and proximal policy optimization (PPO). DPO is a stable and efficient approach to reward-based fine-tuning, while PPO uses reinforcement learning to iteratively improve outputs based on reward signals.

PPO [21] optimizes the model by iteratively interacting with a reward function. It evaluates outputs based on defined metrics, such as accuracy or syntax validity, and adjusts the model to maximize expected rewards. A clipping mechanism in PPO prevents overly large updates to the model parameters, ensuring training stability. However, PPO requires well-defined reward functions, extensive hyperparameter tuning, and significant computational resources, making it complex and resource-intensive for this application.

DPO [22], in contrast, simplifies the process by focusing directly on sampled preferences without requiring explicit reinforcement signals. DPO trains the model to rank outputs based on their quality, as determined by a reward function. This method avoids complex policy adjustments and uses a more straightforward sampling-based approach to refine outputs. DPO requires less computational overhead and delivers more stable results, making it well-suited for tasks such as generating syntactically correct YARA rules. Typically, the training dataset consists of a pair of different text responses given the same query: the chosen response and the rejected one. The chosen response has a higher award score than the rejected response.

In our case, we use DPO due to its simplicity, stability, and reduced resource requirements (see Figure 4). DPO provides a straightforward and interpretable optimization process, making it especially effective in scenarios with limited labeled data and tasks requiring high precision. We define the reward function as:

$$R(y) = \alpha \cdot P(y) + \beta \cdot S(y)$$

where:

- $R(y)$: The reward score for the generated response y
- $P(y)$: A response format validity score (1 if the YARA rule y can be successfully extracted from the response template, 0 otherwise).
- $S(y)$: A binary validity score (1 if the YARA rule y is syntactically valid, 0 otherwise).
- (α, β): Weighting factors to balance the importance of syntax validity and semantic alignment.

We propose an iterative sampling and training algorithm for our YARA rule generation task:

- Step 1: Initialize the model temperature (τ) to encourage diverse responses.
- Step 2: For each CVE in the training set, gather the query data in plain text format.
- Step 3: Use the system prompt and query to generate a response.
- Step 4: Parse the response, extract the YARA rule, and assign a score for the response based on the reward function.
- Step 5: Repeat Steps 3 and 4 five times, leveraging non-zero temperature to explore diverse responses. Retain only the response that has the largest difference in score compared to the response in Step 4.
- Step 6: Form m training pairs using valid and invalid responses by repeating Step 5. Record the number of syntactically incorrect trials in Step 4 as n.
- Step 7: Train the model with these m training pairs, reducing the temperature exponentially based on n.
- Step 8: Repeat the process until τ is zero, with updated τ to refine the model's ability to consistently generate valid YARA rules.

The temperature adjustment in Step 7 follows an exponential decay formula, expressed as:

$$\tau_{i+1} = \tau_i \cdot e^{-\lambda n}$$

where (τ_i) is the current temperature at iteration (i), (λ) is the decay rate constant, and (n) is the number of trials. This ensures that the model progressively focuses on generating more precise outputs as training progresses, making bigger adjustments at the beginning and smaller ones when converging.

D. Iterative Sampling for Rule Matching Quality Improvement

Building upon the syntax correction framework, this step focuses on optimizing the matching quality of YARA rules. Instead of validating syntax alone, the reward mechanism evaluates the effectiveness of the rules in identifying vulnerabilities. The training set includes binaries categorized as containing known CVEs, patched known CVEs, and irrelevant binaries. The reward function for matching quality is defined as:

$$RM(y) = \gamma \cdot R(y) + \delta \cdot F1(y)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

where:

- $RM(y)$: Reward score for the generated response y.
- $R(y)$: The syntax and parsing score from the previous step.
- $F1(y)$: F1 score evaluating the balance between precision and recall when tested on the labeled binary dataset.
- (γ, δ): Weighting factors to balance the importance of target matching and overall F1 score.

The iterative sampling and training algorithm involves:

- Step 1: Initialize the model temperature (τ) to encourage diverse responses.
- Step 2: For each CVE in the training set, gather the query data in plain text format and gather all the testing binaries.
- Step 3: Use the system prompt and query to generate a response.
- Step 4: Parse the response, extract the YARA rule y, and assign the matching score for the response based on the above reward function, by matching the rule y against the known binaries with labels.
- Step 5: Repeat Steps 3 and 4 five times, leveraging non-zero temperature to retrieve different responses. Keep only the response that has the smallest non-zero difference in score compared to the response in Step 4.
- Step 6: Form m training pairs using valid and invalid responses by repeating Step 5. Record the number of syntactically incorrect trials in Step 4 as n.
- Step 7: Train the model with these m training pairs, reducing the temperature (τ) exponentially based on n.
- Step 8: Repeat the process until τ is zero, with updated τ to refine the model's ability to consistently generate valid YARA rules.

This sampling algorithm is similar to the one above for syntax correction, except that the reward score is estimated based on the F1 score, which evaluates the effectiveness of matching the training binaries. Additionally, instead of selecting the pair with the largest score difference as the chosen and rejected responses, we choose the pair with the smallest non-zero difference. This strategy is justified because smaller non-zero differences indicate borderline cases where the model struggles to differentiate quality. Optimizing for such cases helps refine the decision boundary and improves the model's sensitivity to subtle distinctions, ultimately leading to better performance across diverse scenarios.

5. EXPERIMENT EVALUATION

A. Sample Set Building

We start by constructing a CVE vulnerability instance repository containing labeled binaries extracted from well-established firmware images, which serves as a robust testbed for evaluating the system's performance. Our dataset consists of two components: The first comprises popular open-source utility libraries, while the second includes Android OS built-in library vulnerabilities derived from the AOSP dataset [23].

Utility libraries play a significant role in software development, binary analysis, and multimedia processing, but they often exhibit a range of security vulnerabilities. Tools such as addr2line, as, and elfedit, which are respectively used for debugging, assembly, and executable and linkable format (ELF) file manipulation, demonstrate critical flaws across various versions. For instance, addr2line includes vulnerabilities such as CVE-2018-18605, allowing buffer overflows, and CVE-2018-12697, leading to out-of-bounds reads. Similarly, the GNU assembler (as) has been affected by vulnerabilities such as CVE-2017-7230, an integer overflow issue, and CVE-2018-1000019, a stack overflow vulnerability, both of which could enable arbitrary code execution. Multimedia libraries such as ffmpeg, freetype, and libpng also show significant risks, with vulnerabilities such as heap buffer overflows (CVE-2017-7862 in ffmpeg) and use-after-free issues (CVE-2015-8126 in libpng), potentially leading to crashes or remote code execution. These vulnerabilities, arising from issues such as improper input validation and poor memory management, emphasize the need for rigorous security assessments of utility libraries. Table I presents the number of identified vulnerabilities, corresponding library versions, and confirmed CVEs for the open-source utility libraries.

TABLE I: AGGREGATED SUMMARY OF UTILITY LIBRARIES, VERSIONS, AND CONFIRMED CVES

Library	Versions	CVEs	Example CVEs
addr2line	7	72	CVE-2017-14129, CVE-2014-8738, ...
as	2	2	CVE-2017-72.30, ...
elfedit	3	4	CVE-2018-20623, CVE-2017-15996, ...
exif	3	10	CVE-2012-2814, CVE-2012-2840, ...
expat	3	3	CVE-2015-1283, CVE-2012-6702, ...
ffmpeg	45	54	CVE-2017-14059, CVE-2016-7562, ...
freetype	7	63	CVE-2014-9656, CVE-2010-2807, ...
objcopy	2	5	CVE-2018-12699, CVE-2018-12700, ...
objdump	5	16	CVE-2017-8421, CVE-2017-14934, ...
openssl	18	75	CVE-2016-6306, CVE-2015-0289, ...
png	4	6	CVE-2015-8126, CVE-2015-7981, ...
qemu	10	30	CVE-2024-9594, CVE-2024-8612, ...
readelf	2	7	CVE-2017-7209, CVE-2017-9042, ...
sftp	3	3	CVE-2010-4755, CVE-2017-15906, ...
ssh	4	8	CVE-2014-2653, CVE-2011-0539, ...
sshd	7	10	CVE-2016-3115, CVE-2013-4548, ...
tcpdump	3	90	CVE-2017-12902, CVE-2017-13035, ...
xml2	8	38	CVE-2015-8035, CVE-2017-9048, ...

Networking and file-sharing libraries are similarly impacted by security flaws. Tools such as objdump and objcopy contain vulnerabilities such as improper file handling (CVE-2018-6543), which can lead to denial-of-service conditions. Cryptographic libraries, like OpenSSL, suffer from vulnerabilities such as CVE-2016-6306, where improper handling of certificates may result in man-in-the-middle attacks. XML parsing libraries, such as expat and xml2, are also prone to vulnerabilities, including buffer overflows (CVE-2017-9233) and out-of-bounds reads (CVE-2015-8241), which compromise application security. Furthermore, FTP and SSH tools are affected by input handling flaws and directory traversal vulnerabilities, enabling unauthorized access, denial of service, and remote code execution. These widespread vulnerabilities across utility libraries highlight the importance of implementing robust security measures to mitigate evolving threats.

The AOSP dataset [23], hosted on GitHub by Quarkslab, provides a detailed collection of CVEs tailored to the Android operating system. Given Android's extensive integration into various devices, including Internet of Things (IoT) platforms, its security plays a critical role in ensuring device protection. This dataset focuses on vulnerable binary executables, omitting Java-related issues, and allows for an in-depth comparison of pre-patch and post-patch binaries based on source code commit data. With coverage of more than 50 Android system components, such as Media Framework, System, Bluetooth, and SurfaceFlinger, the dataset captures a range of vulnerabilities and their potential effects on device operations. Table II presents the number of vulnerabilities identified in the top 15 built-in libraries of Android system components.

TABLE II: TOP 15 COMPONENTS BY ANDROID OS BUILT-IN LIBRARY VULNERABILITY COUNT

Component	CVEs	High Severity	Critical Severity	Example CVEs
System	202	146	53	CVE-2019-2115
Media Framework	201	106	80	CVE-2019-2176
Mediaserver	136	64	44	CVE-2015-3864
Framework	40	32	5	CVE-2019-2123
libstagefright	21	6	14	CVE-2015-1538
Audioserver	11	9	0	CVE-2017-0418
Libraries	10	6	0	CVE-2016-1839
Bluetooth	8	4	0	CVE-2016-0850
Framework APIs	5	5	0	CVE-2016-3750
system UI	5	5	0	CVE-2017-0638
Binder	4	4	0	CVE-2015-1528
Debuggerd	4	0	2	CVE-2016-2420
Expat	4	1	0	CVE-2012-6702
LibUtils	4	0	4	CVE-2016-3861
OpenSSL & BoringSSL	4	0	2	CVE-2016-0705

Among the 1,000 CVEs, vulnerabilities are categorized by severity and type, including Elevation of Privilege and Remote Code Execution. Key components, such as Media Framework and System account, form a significant portion of the dataset, each containing over 200 vulnerabilities, many of which are high severity. Examples include CVE-2019-2115 in System (a privilege escalation issue), CVE-2019-2176 in

Media Framework (a remote execution risk), and CVE-2015-3864 in MediaServer, which affects media rendering. Other components, such as libstagefright, highlight the risks associated with multimedia processing. This dataset serves as a foundation for understanding vulnerabilities within Android's binaries, aiding efforts to improve the security of IoT devices relying on this architecture. In total, our experiment covers 1,466 vulnerable software records, resulting in 4,218 instances of binary executables for analysis. Additionally, we included 10,000 irrelevant binaries in our experiment to evaluate the false positive rates of the methods.

B. Language Models

This experiment evaluates the performance of five state-of-the-art language models: LLaMA 3.3, Qwen 2, Gemma 2, and Mistral 0.3. These models represent advanced approaches in natural language processing and machine learning, demonstrating varying capabilities in understanding and generating complex patterns from data.

- LLaMA 3.3 [24]: A cutting-edge large language model designed for general-purpose natural language tasks. It focuses on efficiency and scalability, making it suitable for applications requiring high accuracy and rapid inference.
- Qwen 2 [25]: Known for its optimization in handling domain-specific language tasks, Qwen 2 leverages fine-tuned datasets to enhance contextual understanding and generate precise outputs, particularly in technical and specialized areas.
- Gemma 2 [26]: This model excels in multilingual and cross-lingual tasks, offering robust performance across diverse languages. It employs advanced transformer architectures to ensure consistency and coherence in its results.
- Mistral 0.3 [27]: A lightweight yet highly efficient model optimized for resource-constrained environments. Despite its smaller size, Mistral 0.3 delivers competitive performance, making it a practical choice for scalable applications.

The experiment was conducted on a server equipped with a 56-core Xeon Gold 2.3/3.9GHz processor, 100GB of RAM, and two NVIDIA GeForce RTX 6000 cards with (24 GB x 2) of VRAM. The training system was implemented using the HuggingFace Transformer Reinforcement Learning (TRL) library. To assess the performance of the methods, the following evaluation metrics were used:

- Precision: This metric measures the accuracy of positive predictions, indicating the proportion of true positives among all instances predicted as positive. It reflects the system's ability to avoid false alarms.

- Recall: Also known as sensitivity, recall evaluates the system's ability to identify actual positive instances, highlighting how well true positives are detected.
- F1 Score: The F1 score combines precision and recall into a single metric, providing a balanced measure of the system's accuracy, particularly when dealing with imbalanced datasets.

These metrics provide a detailed assessment of the system's performance. We conduct evaluations for each setup, including the original language model, the language model enhanced with syntax correction, and the language model further improved with syntax correction and quality improvement. This layered evaluation helps identify the impact of each enhancement on the model's accuracy and reliability.

TABLE III: PERFORMANCE BENCHMARK

Model (Solution)	Recall	Precision	F1
LLaMA 3.3	0.541	0.019	0.037
Qwen 2	0.622	0.012	0.024
Gemma 2	0.263	0.040	0.069
Mistral 0.3	0.342	0.024	0.045
LLaMA 3.3 (syntax correction)	0.620	0.451	0.522
Qwen 2 (syntax correction)	0.774	0.672	0.719
Gemma 2 (syntax correction)	0.561	0.836	0.671
Mistral 0.3 (syntax correction)	0.832	0.681	0.749
LLaMA 3.3 (syntax correction + quality improvement)	0.992	0.781	0.874
Qwen 2 (syntax correction + quality improvement)	0.986	0.893	0.937
Gemma 2 (syntax correction + quality improvement)	0.912	0.866	0.888
Mistral 0.3 (syntax correction + quality improvement)	0.956	0.901	0.928

C. Experimental Results

Table III shows the experimental results. The evaluation of the language models in their original configurations highlights their limited ability to handle the task effectively. Metrics for recall, precision, and F1 score remain low across all models. For example, LLaMA 3.3 achieves a recall of 0.541 but a precision of only 0.019, resulting in an F1 score of 0.037. Qwen 2, Gemma 2, and Mistral 0.3 show similar patterns, with F1 scores ranging from 0.024 to 0.069. These outcomes suggest that the models, in their initial states, struggle to balance the identification of true positives with the minimization of false positives.

When syntax correction is applied, significant improvements are observed in all models. LLaMA 3.3, for instance, achieves an F1 score of 0.522, a marked improvement from its original performance. Qwen 2, Gemma 2, and Mistral 0.3 also show improved metrics, with F1 scores increasing to 0.719, 0.671, and 0.749, respectively. Syntax correction addresses structural inconsistencies, enabling the models to better interpret and process data. This adjustment results in higher recall and precision, demonstrating its impact on performance.

The combination of syntax correction and quality improvement delivers the best results across all models. For example, LLaMA 3.3 reaches a recall of 0.992 and an F1 score of 0.874, while Qwen 2 achieves a recall of 0.986 and an F1 score of 0.937. Mistral 0.3 and Gemma 2 show similar improvements, with F1 scores of 0.928 and 0.888, respectively. These enhancements refine both the input and the underlying understanding of the models, leading to improved predictions and reduced errors. This approach highlights the benefits of combining structural corrections with quality refinements to achieve optimal performance.

6. CONCLUSION

The proposed method significantly improves vulnerability patch verification by combining large language models, syntax correction, and quality enhancement techniques. Experimental findings demonstrate the limitations of initial models, and the performance gains achieved through structured refinement. By automating the generation of YARA rules and focusing on syntax and matching precision, the system overcomes challenges faced by traditional version-based detection methods. This approach offers a scalable and accurate solution for verifying software vulnerabilities, addressing the demands of environments requiring high security and reliability.

7. FUTURE WORK

While our experiments show promising results in model training across a subset of tasks and datasets, new avenues of research remain open for further exploration and improvement. We plan to benchmark our model's performance on each dataset separately and address emerging common weakness enumerations (CWEs). While our study focuses on the existing architecture, many new transformer-based and other neural network variants are emerging. Further research could involve fine-tuning these novel architectures under similar conditions to benchmark performance, parameter efficiency, and speed. By pursuing these directions, we hope to deepen our understanding of model fine-tuning, leading to improved vulnerability patch verification.

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] R. Ramachandran. "Qualys Top 20 Most Exploited Vulnerabilities." 2003. Accessed: Jan. 8, 2025. [Online]. Available: https://blog.qualys.com/vulnerabilities-threatresearch/2023/09/04/qualys-top-20-exploited-vulnerabilities

[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] H. Ghanbari, K. Koskinen, and Y. Wei, "From SolarWinds to Kaseya: The rise of supply chain attacks in a digital world," *J. Inf. Technol. Teach. Cases*, Nov. 2024, doi: 10.1177/20438869241299823.

[5] P. Ladisa, H. Plate, M. Martinez, and O. Barais, "SoK: Taxonomy of attacks on open-source software supply chains," in *Proc. IEEE Symp. Secur. Priv. (SP)*, San Francisco, CA, USA, May 2023, pp. 1509–1526.

[6] Y. Shen, X. Gao, H. Sun, and Y. Guo, "Understanding vulnerabilities in software supply chains," *Empir. Softw. Eng.*, vol. 30, no. 20, Nov. 2024.

[7] Fraunhofer FKIE. "GitHub—fkie-cad/FACT_core: Firmware analysis and comparison tool." Accessed: Jan. 8, 2025. [Online]. Available: https://github.com/fkie-cad/FACT_core

[8] E-M-B-A. "GitHub—e-m-b-a/embark: EMBArk—The firmware security scanning environment." Accessed: Jan. 8, 2025. [Online]. Available: https://github.com/e-m-b-a/embark

[9] Intel. "GitHub—intel/cve-bin-tool: The CVE binary tool helps you determine if your system includes known vulnerabilities." Accessed: Jan. 8, 2025. [Online]. Available: https://github.com/intel/cve-bin-tool

[10] M. Beninger, P. Charland, S. H. H. Ding, and B. C. M. Fung, "ERS0: Enhancing military cybersecurity with AI-driven SBOM for firmware vulnerability detection and asset management," in *Proc. 16th Int. Conf. Cyber Conflict: Over the Horizon (CyCon)*, Tallinn, Estonia, May 2024, pp. 141–160.

[11] VirusTotal. "GitHub—VirusTotal/yara: The pattern matching Swiss knife." Accessed: Jan. 8, 2025. [Online]. Available: https://github.com/VirusTotal/yara

[12] J. Clause, W. Li, and A. Orso, "Dytan: A generic dynamic taint analysis framework," in *Proc. Int. Symp. Softw. Test. Anal. (ISSTA)*, London, U.K., July 2007, pp. 196–206.

[13] L. Li, S. H. H. Ding, Y. Tian, B. C. M. Fung, P. Charland, W. Ou, L. Song, and C. Chen, "VulANalyzeR: Explainable binary vulnerability detection with multi-task learning and attentional graph convolution," *ACM Trans. Privacy Secur.*, vol. 26, no. 3, art. no. 28, pp. 1–25, Apr. 2023.

[14] A. Fioraldi, D. Maier, H. Eißfeldt, and M. Heuse, "AFL++: Combining incremental steps of fuzzing research," in *Proc. 14th USENIX Workshop on Offensive Technologies (WOOT)*, Aug. 2020.

[15] F. Wang and Y. Shoshitaishvili, "Angr—The next generation of binary analysis," in *Proc. IEEE Cybersecur. Dev. (SecDev)*, Cambridge, MA, USA, Sept. 2017, pp. 8–9.

[16] S. H. H. Ding, B. C. M. Fung, and P. Charland, "Asm2Vec: Boosting static representation robustness for binary clone search against code obfuscation and compiler optimization," in *Proc. IEEE Symp. Secur. Priv. (SP)*, San Francisco, CA, USA, May 2019, pp. 472–489.

[17] 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*.

[18] L. Giray, "Prompt engineering with ChatGPT: A guide for academic writers," *Ann. Biomed. Eng.*, vol. 51, no. 12, pp. 2629–2633, Dec. 2023.

[19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.

[20] Z. Yao, C. Li, X. Wu, S. Youn, and Y. He, "A comprehensive study on post-training quantization for large language models," 2023, *arXiv:2303.08302*.

[21] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” 2017, *arXiv:1707.06347*.

[22] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, “Direct preference optimization: Your language model is secretly a reward model,” in *Proc. 37th Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, New Orleans, LA, USA, Dec. 2023, pp. 53728–53741.

[23] A. Challande, R. David, and G. Renault, “Building a commit-level dataset of real-world vulnerabilities,” in *Proc. 12th ACM Conf. Data Appl. Secur. Privacy (CODASPY)*, Baltimore, MD, USA, Apr. 2022, pp. 101–106.

[24] A. Grattafiori et al., “The Llama 3 herd of models,” 2024, *arXiv:2407.21783*.

[25] J. Bai et al., “Qwen technical report,” 2023, *arXiv:2309.16609*.

[26] M. Riviere et al., “Gemma 2: Improving open language models at a practical size,” 2024, *arXiv:2408.00118*.

[27] A. Q. Jiang et al., “Mistral 7B,” 2023, *arXiv:2310.06825*.

CyCon 2025: The Next Step
17th International Conference on Cyber Conflict
C. Kwan, N. Gratzer, K. Podiņš, M. Tolppa (Eds.)
2025 © NATO CCDCOE Publications, Tallinn

Permission to make digital or hard copies of this publication for internal use within NATO and for personal or educational use when for non-profit or non-commercial purposes is granted provided that copies bear this notice and a full citation on the first page. Any other reproduction or transmission requires prior written permission by NATO CCDCOE.

Next Steps in Cyber Blue Team Automation—Leveraging the Power of LLMs

Allard Dijk
Netherlands Defence Academy
Den Helder, The Netherlands
ad.dijk@mindef.nl

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

Cosimo Melella
NATO Cooperative Cyber Defence Centre of Excellence
Tallinn, Estonia
cosimo.melella@ccdcoe.org

Mauno Pihelgas
Tallinn University of Technology
Tallinn, Estonia
maunopihelgas@gmail.com

Risto Vaarandi
Tallinn University of Technology
Tallinn, Estonia
risto.vaarandi@taltech.ee

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

Abstract: In 2021, driven by the ongoing advancements in artificial intelligence (AI) and automation, previous works [1], [2] introduced architectures for fully automated blue teams in cyber defense exercises such as Locked Shields (LS). Since then, technological and scientific progress has further accelerated. In particular, the rapid evolution of generative AI through large language models (LLMs) has significantly enhanced the capabilities of cybersecurity automation.

This paper reviews how cyber blue team automation can benefit from these recent advances, with a focus on how generative AI and LLMs are reshaping automation strategies for defending complex cyber infrastructure. Using the LS exercise as a case study, we discuss how generative AI-based automation can address the growing complexity of cyber threats. Our paper presents promising directions on how generative AI can enhance fully automated blue teams, and it addresses a major research gap—

the lack of high-quality datasets for training and evaluation in this field. To address this challenge, we introduce a novel dataset containing labeled network traffic and end-host logs, collected during the “partners’ run” preceding LS 2024. This dataset is derived from over 400 GB of captured network traffic and more than 6 million log entries. It captures real-world red team behavior and is made publicly available to foster research and AI development in the field of blue team automation.

We conclude with future research challenges in automated cyber defense.

Keywords: *automated cyber defense, Locked Shields, artificial intelligence, large language models, dataset*

1. INTRODUCTION

Artificial intelligence (AI) is disrupting almost every field at an unprecedented pace. This also includes the field of cybersecurity, where AI is transforming both the attack and defense landscapes. While attackers increasingly exploit AI to automate and enhance their cyber exploitation methods, defenders are leveraging AI to improve detection, response, and mitigation strategies. For example, AI can identify suspicious patterns and anomalies in network traffic or application logs, pinpointing potential threats faster and with greater accuracy than traditional methods [3]. Beyond detection, AI is increasingly being used in response automation, such as orchestrating defense mechanisms or remediating vulnerabilities with or without human intervention [4], [5].

However, even though AI has made significant progress, it is not yet advanced enough to fully replace human experts in cyber defense. For instance, AI struggles to adapt to scenarios that deviate from its training data [6]. Furthermore, the potential for false positives and the lack of high-quality labeled data limits the effectiveness of AI in real-world applications. Building on this, Zhang et al. [7] explore the applications of AI in cybersecurity, including user access authentication, network situation awareness, dangerous behavior monitoring, and abnormal traffic identification. They emphasize the role of AI in enhancing cybersecurity measures and propose a conceptual human-in-the-loop cybersecurity model, stressing the importance of human involvement alongside AI systems.

In this paper, we analyze the current capabilities of AI in the context of automating cyber defense. We focus on the use of such automation in live-fire cyber defense exercises such as Locked Shields (LS), because these exercises provide an ideal testing ground for new technology. Starting with a framework that Meier et al. developed in 2021 [1], we discuss the impact of AI developments that have happened since then and we present the next steps toward the vision of a fully automated defense team at a cyber defense exercise. We base our discussion on ongoing research efforts suitable for LS, the world's largest international live-fire cyber defense exercise. We also publish a labeled dataset from this exercise in order to allow the research community to develop and test their models with realistic data and potentially use it to train or improve LLMs for cybersecurity automation.

In summary, the main contributions of our paper are:

- A retrospective of the latest developments in the context of generative AI and how they affect blue team automation (Section 3);
- A discussion of the main use cases for generative AI for blue team automation (Section 4);
- A plan for the next steps toward the vision of an automated blue team, and the challenges and opportunities that generative AI brings (Section 5);
- A novel dataset containing labeled network traffic and end-host logs to foster research (including training of new LLMs) (Section 6).

2. BACKGROUND ON CYBER DEFENSE EXERCISES

Cyber defense exercises are critical for enhancing operational readiness, fostering interdisciplinary cooperation, and improving cyber defenses in the ever-evolving cyber domain. Among the most prominent examples is LS, an annual live-fire exercise organized by the NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) since 2010 [8]. It has recently gained additional relevance as a testbed for incorporating AI into cyber defense operations.

LS is a two-day, defense-oriented exercise centered around a fictional geopolitical conflict. Blue teams (BTs), composed of rapid-reaction cybersecurity units, are tasked with defending the IT and critical infrastructure of the fictional country Berylia against the red team (RT), which represents a hostile state, Crimsonia. At a high level, the tasks of the BTs can be grouped into four stages: initial hardening (harden systems before the attacks start), monitoring and response (detect and mitigate attacks), reporting (document the observed attacks), and recovery (restore gamenet systems from backups or with help of the exercise organizers).

The BTs are the main training audience in LS, and they are scored across various categories, including defending against RT attacks, incident reporting, and maintaining service availability. Each BT is responsible for maintaining the uptime and security of over 140 physical and virtual hosts, which include standard IT systems, industrial control systems, and specialized components such as 5G infrastructure.

Recent research [1], [9] has introduced AI into LS, showcasing its potential to enhance defense strategies. Such AI-powered systems can improve defense capabilities by offering faster threat detection and response, scalability to manage complex infrastructures, and continuous learning from attack patterns.

3. THE VISION OF AN AUTOMATED BLUE TEAM

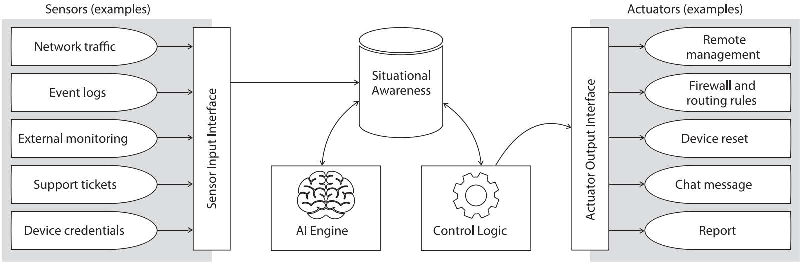
In 2021, Meier et al. developed a general architecture for an AI-powered player in cyber defense exercises [1]. The architecture is depicted in Figure 1. It consists of the following main components:

Sensors are components that provide measurements or data. Examples of sensors and the data that they provide include: network traffic, event logs, device credentials, or support tickets by users.

Actuators are components that perform actions in the gamenet. Examples of actuators include: remote management (e.g., via SSH or RDP), modifications of firewall rules, reset or reboot of a device, or generating a response to a support ticket.

In between the sensors and the actuators are three additional building blocks: the *situational awareness database* (contains all sensor data), the *AI engine* (learns and applies AI models in order to enhance the situational awareness database), and the *control logic* (triggers actuators depending on the contents of the situational awareness database).

FIGURE 1: ARCHITECTURE FOR AN AUTOMATED BLUE TEAM, DEVELOPED BY MEIER ET AL. [1]



In 2021, the authors of [1] could not foresee the upcoming generative AI revolution, most notably marked by the release of ChatGPT in November 2022. This release marked a milestone in the evolution of generative AI, demonstrating its ability to engage in complex, human-like conversations and solve problems.

Fundamentally, generative AI is a type of artificial intelligence designed to create content rather than simply analyze or classify existing data. Generative AI models (e.g., OpenAI's ChatGPT,[1] Google's Gemini,[2] or Meta's Llama[3]) can produce text, images, code, and other creative outputs based on patterns they have learned during training. In the case of LLMs, the focus is on generating coherent, context-aware text that mimics human language.

At their core, LLMs are built on a neural network architecture called Transformers, which excels at processing and generating sequential data, such as language. These models are trained on massive datasets, including books, articles, websites, and other text sources, to identify statistical relationships between words, phrases, and contexts. The goal is not to "understand" language as humans do but to generate text that aligns with patterns and structures found in natural language.

Today's LLMs can produce high-quality outputs and handle a wide range of tasks across industries, including IT and cybersecurity. There, LLMs have proven to be a valuable assistant in areas such as debugging code [10], finding vulnerabilities [11] and analyzing system logs [12]. However, it is important to note that LLMs lack true comprehension or reasoning. Instead, they generate content based solely on learned patterns and can produce biased or incorrect information if such issues exist in its training data.

4. APPLICATIONS OF LLMS IN BLUE TEAM AUTOMATION

As an AI technology, LLMs primarily influence the "AI engine" component in Figure 1. However, they significantly expand the possibilities for processing sensor data and generating inputs for actuators. In this section, we explore the key use cases where LLMs offer notable advantages over previously available technologies.

We categorize these use cases according to the four stages of a cyber defense exercise: initial hardening, monitoring and response, reporting, and recovery (see Section 2). Table I provides an overview, and the remainder of this section explains all use cases in more detail.

[1] https://chatgpt.com/
[2] https://gemini.google.com/
[3] https://llama.com/

TABLE I: OVERVIEW OF USE CASES WHERE LLMS PROVIDE A SIGNIFICANT ADVANTAGE COMPARED TO PREVIOUS METHODS

Stage	Use Cases for LLMs
Initial hardening	• Identification and fixing of vulnerabilities and misconfigurations in software (Section 4.A)
Monitoring and response	• Analyzing network traffic for malicious activities (Section 4.B) • Analyzing event logs for malicious activities (Section 4.C) • Parse support tickets, trigger corresponding actions, and generate responses (Section 4.D) • Generate commands and configurations for remote management (Section 4.E)
Reporting	• Link incidents to IoCs and generate human-readable reports (Section 4.F) • Generate human-readable reports required for the exercise (e.g., post-incident summaries) (Section 4.E)
Recovery	• Identifying and documenting affected systems for recovery prioritization (Section 4.C) • Reverting devices, misconfigurations, and patch failures using rollback mechanisms such as backups and snapshots (Section 4.E) • Generating comprehensive post-recovery analysis and lessons learned documentation (Section 4.F)

A. Detecting Vulnerabilities and Misconfigurations

Software vulnerabilities are flaws or weaknesses in an application's design, implementation, or configuration. In an exercise like LS, these weaknesses can include anything from poorly secured web applications and misconfigured Docker containers to hidden backdoors intentionally placed by the RT. For a BT, discovering and remediating these vulnerabilities quickly is vital.

Traditional methods have proven effective but are often time-consuming and demand specialized expertise. For example, traditional static analysis techniques (see [13]) have uncovered numerous bugs at scale, but they struggle to keep pace with increasingly complex systems. Similarly, dynamic taint analysis [14] set early precedents for automated exploit detection but faces scalability issues in modern environments.

LLM-based approaches offer greater flexibility and efficiency. Systems like LProtector, built on GPT-based models, excel at detecting vulnerabilities in large codebases [15]. By training on extensive code repositories, these models can identify issues like SQL injection, remote code execution, and cross-site scripting with remarkable accuracy [16], [17]. At the same time, the use of AI-driven code generation tools (e.g., GitHub Copilot) has been scrutinized for potential security risks [18].

LLMs can similarly detect misconfigurations by examining database or web server settings, pinpointing insecure network parameters or permissive access controls [19]. This proactive approach helps preempt exploitation by simulating possible attack vectors.

Research also suggests that AI-driven rule adjustments for security policies can keep pace with emerging threats [20]. By prioritizing vulnerabilities according to severity, defenders can allocate resources more effectively [15]. Finally, while direct LLM-based remediation remains an emerging topic, previous efforts in machine-learning-driven security automation indicate a promising direction [21].

B. Network Traffic Analysis

With the integration of data-mining-based algorithms and, more recently, LLMs, the field of network traffic analysis has seen significant advancements. Traditional algorithms such as decision trees or support vector machines were employed to analyze traffic for detecting patterns and anomalies [22], [23].

LLM-based approaches introduced a new paradigm in traffic analysis: LLMs can process and understand vast amounts of unstructured data (e.g., network logs) and automate incident response actions [24]. They can recommend or autonomously execute predefined responses to threats, reducing the time and effort required from human analysts.

LLM-based methods can also classify different malware types with limited amounts of training data compared to state-of-the-art methods: Even though the structure of network protocols is different from natural language, Stein et al. [25] demonstrate that transformer-based models can capture and learn the intricate sequential patterns. Unlike many LLMs that are pre-trained and then fine-tuned, RTIDS [26] shows that a transformer-based intrusion detection system (IDS) can achieve promising results when trained from scratch by batching collections of network flows during the training process. However, such supervised approaches require extensive labeled datasets, which can be challenging to obtain. The LSPR23 [9] dataset and the LSPR24 dataset released with this paper facilitate this kind of research.

C. Log Analysis

In the last two decades, several data-mining-based algorithms have been proposed for analyzing textual event logs (cf. [27]). These algorithms focus on the detection of line patterns or templates from textual event logs, with the following example illustrating two example log messages and a template representing them:

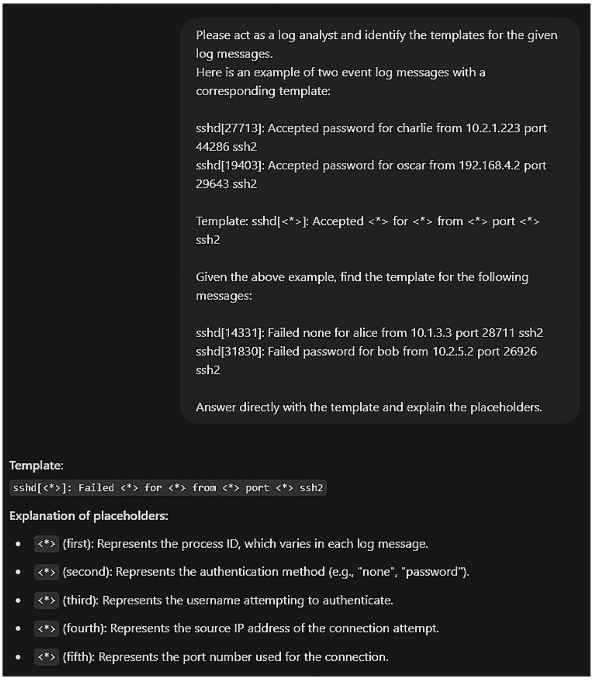
```
sshd[27713]: Accepted password for charlie from 10.2.1.223 port 44286 ssh2
sshd[19403]: Accepted password for oscar from 192.168.4.2 port 29643 ssh2
Template: sshd[<*>]: Accepted <*> for <*> from <*> port <*> ssh2
```

The purpose of the template detection is to identify variable parts in event logs and replace these parts with a wildcard (<*> in the above example). This is useful for detecting event types from logs and assisting the development of event parsing rules. Additionally, these insights can support the identification and documentation of affected systems in a cyber exercise, enabling prioritization in recovery efforts by understanding the scope and impact of incidents.

Recent studies have demonstrated the potential of using LLMs for template detection tasks [28], [29], [30], [31], [32], [33]. Although LLM-based approaches require more computational resources than traditional data-mining-based algorithms and so tend to be slower [28], [32], [33], they have several benefits. For example, some LLMs have the ability to infer a correct template even from insufficient log data [33].

Some algorithms like LLMParser [29] and LogPPT [31] use *fine-tuning* of local LLMs, which involves additional training of LLMs with examples of event log messages and expected templates. The other and more commonly used approach is *in-context learning*, which involves providing an LLM with instructions (*prompt*) on the template detection task [28], [30], [32], [33]. Usually, the prompt contains some examples of event log messages with expected template(s) and the actual event log messages. Figure 2 shows an example using ChatGPT.

FIGURE 2: LLM PROMPT TO EXTRACT A TEMPLATE FROM LOG MESSAGES (USING CHATGPT 4o)



Please act as a log analyst and identify the templates for the given log messages.
Here is an example of two event log messages with a corresponding template:

```
sshd[27713]: Accepted password for charlie from 10.2.1.223 port 44286 ssh2
sshd[19403]: Accepted password for oscar from 192.168.4.2 port 29643 ssh2
```

Template: sshd[<*>]: Accepted <*> for <*> from <*> port <*> ssh2

Given the above example, find the template for the following messages:

```
sshd[14331]: Failed none for alice from 10.1.3.3 port 28711 ssh2
sshd[31830]: Failed password for bob from 10.2.5.2 port 26926 ssh2
```

Answer directly with the template and explain the placeholders.

Template:

```
sshd[<*>]: Failed <*> for <*> from <*> port <*> ssh2
```

Explanation of placeholders:

- `<*>` (first): Represents the process ID, which varies in each log message.
- `<*>` (second): Represents the authentication method (e.g., "none", "password").
- `<*>` (third): Represents the username attempting to authenticate.
- `<*>` (fourth): Represents the source IP address of the connection attempt.
- `<*>` (fifth): Represents the port number used for the connection.

Since the response from an LLM is provided in natural language, the algorithms that use LLMs through in-context learning must parse the LLM's answer in order to identify the templates in the received response.

Algorithms that rely on in-context learning can be supervised or unsupervised. Existing supervised algorithms LILAC [28] and DivLog [30] assume that the human expert has to create a larger set of example log messages with a correct template provided for each message. When constructing the prompt, the algorithms analyze the event log messages supplied by the user and select the most appropriate examples from the set prepared by the human expert. The main drawback of supervised algorithms is the need for such data sets with expert-supplied templates. LUNAR [32] and LLM-TD [33] are unsupervised algorithms which do not employ large, manually created example sets to build prompts but rather use prompts with static instructions and examples. LLM-TD mines *syslog* messages, whereas LUNAR employs a hierarchical clustering algorithm to detect similar messages that are suitable for submitting to the LLM in one query.

From the aforementioned algorithms that rely on in-context learning, DivLog, LILAC, and LUNAR employ public LLMs (e.g., ChatGPT through the OpenAI interface). Since LLM-TD has been specifically designed for analyzing security event logs, it uses local LLMs through the Ollama framework in order to avoid submitting potentially sensitive log data to external service providers.

D. Interacting with Humans

In a cyber defense exercise, BTs typically handle a continuous influx of user inquiries, status updates, and incident reports. Traditionally, these tasks were allocated to human analysts who had to parse support tickets and either execute relevant technical actions or delegate issues to other specialized team members.

LLMs bring efficiency to this process by interpreting the natural-language content of support tickets, extracting critical information (e.g., IP addresses, error codes, account names), and aligning them with the corresponding technical actions [34]. For instance, an LLM-based system may scan a high volume of tickets, identify distinct categories such as "hardware failures" or "phishing suspicions," and automatically open an internal task for resetting credentials or blocking malicious domains [35]. Thereby, LLMs significantly compress the review cycle time [36].

LLMs can also create human-readable summaries and incident reports. Rather than manually drafting a lengthy post-incident description, analysts can rely on the LLM to compile system logs, relevant indicators of compromise (IoCs), and incident timelines into a coherent narrative [37]. In exercises that score teams on thorough and timely incident reporting, this functionality ensures both clarity and consistency, thereby reducing the risk of miscommunication [38].

E. Remote Management

Beyond assisting human interactions, LLMs also play a pivotal role in controlling the infrastructure directly. IT environments require configuration files, scripts, or remediation commands to be maintained in real time [39]. Handling this efficiently can be challenging, especially under the time pressure of a live-fire exercise where multiple systems need simultaneous updates or patches [40].

LLMs can convert high-level policy descriptions or abstract instructions into code or commands, enabling the automated generation of restoration scripts and configurations needed to recover compromised or misconfigured systems [41], [42]. For instance, when the control logic component flags an unauthorized process on a critical server, the LLM can propose a suitable script to terminate that process, quarantine files, or modify a firewall configuration [43]. This eliminates the need for a human operator to research appropriate syntax or recall rarely used commands. In addition, by integrating with version control repositories, an LLM can track system configurations over time, offering automated rollbacks if an action inadvertently disrupts legitimate services [42].

A particularly promising avenue involves coupling LLMs with “computer use” modes, where the LLM can directly interface with network devices or cloud-based management consoles. In this scenario, the language model constructs the commands, verifies them against known best practices or policy constraints, and then executes them autonomously or with minimal supervision [44]. While this streamlines remote management, it also raises questions about access controls and the risk of an attacker manipulating the LLM to issue malicious commands [45].

F. Integrating Threat Intelligence Feeds and SIEM Systems

Leveraging external threat intelligence feeds, such as those provided by the Malware Information Sharing Platform (MISP) [46], is critical for enhancing cybersecurity workflows by enabling the sharing of IoCs and fostering collaboration [47]. LLMs offer a transformative approach to processing and integrating this data into security information and event management (SIEM) systems by automating tasks like ingestion, contextualization, and prioritization [48].

In scenarios like LS, LLMs could dynamically analyze threat intelligence feeds, categorize threats by severity, and link related IoCs to broader campaigns, providing actionable insights that improve situational awareness and decision-making [49]. Integrating external intelligence feeds with SIEMs through LLMs creates a pipeline for correlating IoCs with internal logs, ranking threats by relevance, augmenting data with contextual analysis, and suggesting automated responses, such as blocking IPs or quarantining devices [50].

This synergy reduces the load on analysts, enhances detection speed, and facilitates post-event analysis by generating comprehensive reports. Furthermore, it supports the creation of detailed post-recovery analysis and lessons learned documentation, ensuring organizations can refine their defensive measures based on past incidents [51]. Despite these advantages, challenges include ensuring data quality, maintaining privacy through locally hosted or fine-tuned models, and addressing interpretability issues in LLM outputs to justify their decisions [52]. Experimental frameworks combining MISP, SIEMs, and LLMs could provide valuable insights into real-world applications, paving the way for more efficient and automated cyber defense [53].

5. CHALLENGES AND NEXT STEPS

Based on the insights from the previous sections, we now discuss the current challenges at a higher level and outline the next steps toward the vision of an automated BT.

A. Challenges

Data availability: Training or fine-tuning LLMs requires extensive and high-quality datasets. Without sufficient data, models may underperform or fail to generalize effectively. Cyber defense exercises such as LS provide a good basis for gathering high-quality training data. However, it is also important to be aware of the differences between exercises and real-world incidents where attacks are more subtle and leverage a larger set of strategies.

Prompt engineering: Well-crafted prompts are critical for guiding LLM behavior. This is especially challenging because the ultimate vision is for these prompts to be generated automatically, without human involvement. A related challenge is the so-called context size of an LLM. This refers to the maximum amount of information it can process at once (i.e., its capacity to "remember"). If an LLM needs to process vast amounts of information (e.g., log files or network data), preprocessing is required to provide only the LLM the relevant information.

Hallucination: Hallucination of LLMs refers to their tendency to generate inaccurate or fabricated information and is difficult to identify. This occurs in any LLM application, but in the context of a fully automated system, it can have greater consequences because there is no "human in the loop" who could detect the hallucination.

Integration complexity: Implementing seamless interfaces between the various components (see Figure 1) is a significant engineering challenge.

Computational power: Running LLMs can demand substantial computational resources. However, there are promising alternatives, such as models optimized for commercial off-the-shelf GPUs, and cloud-based models that can mitigate this issue.

Measuring effectiveness: Assessing the performance of an automated BT (and its components, including the LLMs) in a reproducible manner is critical. While cyber defense exercises like LS provide a valuable opportunity for such experiments, the fact that these exercises typically happen only once a year slows progress. Ideally, there should be a reproducible environment for testing systems multiple times per year.

B. Next Steps

To integrate LLMs into an automated BT, we propose the following next steps:

Integrate and evaluate individual LLM components: We estimate that the following use cases have the highest potential for LLMs to achieve a significant advantage compared to traditional solutions. Therefore, they should be addressed first:

- Automating support ticket processing: Utilize LLMs to convert human-written support tickets into actionable technical instructions, such as code, commands, or configuration files.
- Generating human-readable reports: Leverage LLMs to create detailed, easily understandable reports or responses to support tickets.
- Detecting and fixing misconfigurations: Use LLMs to identify system misconfigurations and generate precise corrective actions, including code or commands.
- Combining data for actionable insights: Employ LLMs to analyze and synthesize data from multiple sources, such as event logs and network traffic, to uncover valuable insights and patterns.

Establish a reproducible testing environment: To allow for consistent evaluation and improvement of the automated BT, there needs to be a testing environment that supports frequent, repeatable tests. To maximize efficiency, the testing environment should operate without requiring manual actions from human experts (e.g., from an RT), as such resources are often difficult to obtain. Instead, the environment could leverage automated scenarios, potentially running within a cyber range to simulate realistic attack-defense-interactions. Furthermore, RT automation is a related field of research that we did not cover in this paper. However, such a testing environment could serve as a playground for experimenting with automated RTs versus automated BTs, fostering advancements in both areas and enabling comprehensive evaluations of emerging defensive and offensive strategies.

6. THE LSPR24 DATASET

Collected during the partners' run prior to Locked Shields 2024, the LSPR24 dataset provides a solid foundation for BT automation research. We publish it to facilitate AI-driven model training by the research community [54]. The dataset also enables validation of automated frameworks that integrate logs from multiple sources, and its structure allows for more effective log analysis and automated responses, particularly in combination with LLMs.

A key feature of LSPR24 is that it originates from a complex, realistic environment. Spanning over 400 GB of captured network traffic, it represents diverse hardware configurations, software stacks, and user behaviors, making it a robust resource for machine learning. Host logs, network flows, and Suricata/Zeek outputs help researchers observe both benign and malicious behaviors, including lateral movement and command-and-control (C2) methods.

Figure 3 presents the LSPR24 high-level network map for LSPR24, connecting the government, military, and energy sectors. It integrates advanced technologies like 5G, AI surveillance, and hybrid-cloud systems with traditional satellite communication, air defense, and border security.

FIGURE 3: HIGH-LEVEL OVERVIEW OF THE GAMENET IN WHICH WE CAPTURED THE LSPR24 DATASET

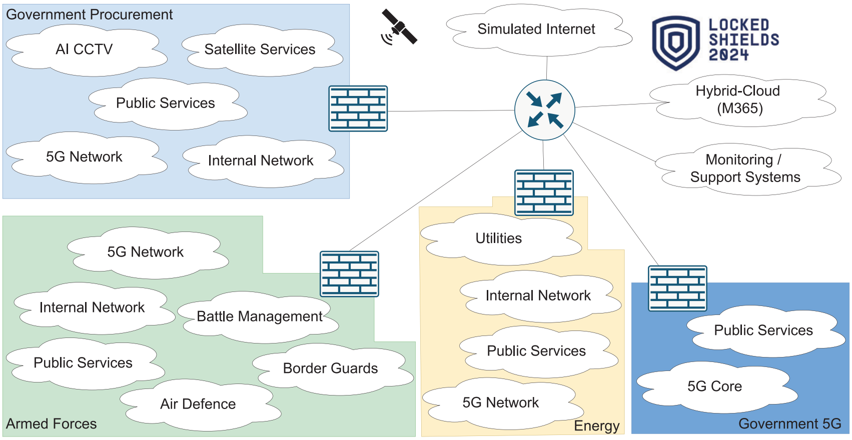


FIGURE 4: HOURLY ACTIVITIES IN THE LSPR24 DATASET

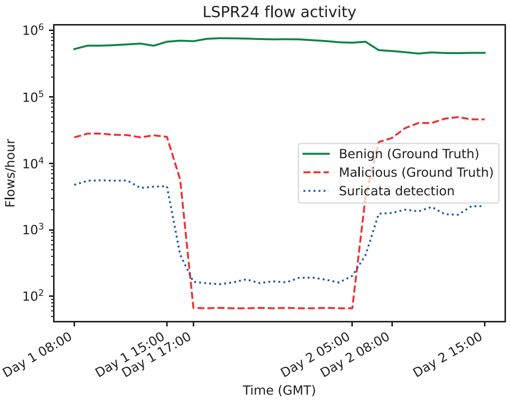


Figure 4 shows the flow activity throughout of the dataset. Benign traffic (green) remains consistently high—around 1 million flows per hour—while malicious traffic (red) shows more fluctuation. Notably, it dips sharply around Day 1 at 17:00 GMT, then intensifies again on Day 2. The detections of Suricata, a popular traditional intrusion-detection system, (blue) show many false positives (during times when there is no malicious activity). This indicates that more sophisticated technology is needed to detect attacks.

LSPR24 contains 20 million flows, two billion packets, and 287 GB of transferred data over 31.6 hours. The collection spans activity across 13,000 IPv4 and IPv6 addresses, including 372 linked to the RT.

Compared to its predecessor, LSPR23 [9], LSPR24 addresses gaps in IDS signatures, including those targeting Cobalt Strike beacon traffic. It also improves internal flow labeling to accurately classify stepping-stone attacks, enhancing the analysis of suspicious behavior within a defended network.

7. CONCLUSION

This paper revisited the vision of a fully automated blue team, originally published in 2021, and explored how advancements in generative AI can contribute to realizing this vision. By examining the potential applications of generative AI in cyber defense, we identified both opportunities and challenges that remain in the field.

A key practical obstacle is the availability of high-quality datasets necessary for the development and evaluation of AI models. To address this gap and foster further research, we published a new labeled dataset comprising network flows and event logs collected during Locked Shields, the world's largest live-fire cyber defense exercise.

By providing insights into generative AI's role and offering resources to the research community, this paper serves as a foundation and guideline for advancing toward the goal of a fully automated blue team. Future research should focus on addressing the highlighted challenges and building on the resources provided to achieve this vision.

Notably, generative AI is useful not only for blue teams but also for red teams, potentially creating a new dynamic between increasingly automated adversarial and defensive systems [55]. While this paper focused on the blue team perspective, the interplay between blue team and red team automation creates another relevant research direction for the future.

ACKNOWLEDGMENTS

Risto Vaarandi's work was supported by the Estonian Centre of Excellence in Artificial Intelligence (EXAI), funded by the Estonian Ministry of Education and Research grant TK213.

REFERENCES

[1] R. Meier, A. Lavrenovs, K. Heinäaro, L. Gambazzi, and V. Lenders, "Towards an AI-powered player in cyber defence exercises," in *Proc. 13th Int. Conf. Cyber Conflict (CyCon)*, Tallinn, Estonia, May 2021.

[2] A. Kott, Ed., *Autonomous Intelligent Cyber Defense Agent (AICA): A Comprehensive Guide, , Advances in Information Security*, vol. 87. Cham: Springer, 2023.

[3] I. H. Sarker, M. H. Furhad, and R. Nowrozy, "AI-driven cybersecurity: An overview, security intelligence modeling and research directions," *SN Comput. Sci.*, vol. 2, Mar. 2021.

[4] S. Lysenko, "The role of artificial intelligence in cybersecurity: Automation of protection and detection of threats," *Econ. Aff.*, vol. 69, Feb. 2024.

[5] L. Alevizos, "Automated cybersecurity compliance and threat response using AI, blockchain and smart contracts," *Int. J. Inf. Technol.*, Dec. 2024.

[6] L. Gehri, R. Meier, D. Hulliger, and V. Lenders, "Towards generalizing machine learning models to detect command and control attack traffic," in *Proc. 15th Int. Conf. Cyber Conflict: Meeting Reality (CyCon)*, Tallinn, Estonia, May 2023.

[7] Z. Zhang et al., "Artificial intelligence in cyber security: Research advances, challenges, and opportunities," *Artif. Intell. Rev.*, vol. 55, Feb. 2022.

[8] NATO Cooperative Cyber Defence Centre of Excellence, "NATO Locked Shields." Accessed: Jan. 4, 2025. [Online]. Available: https://ccdcoe.org/exercises/locked-shields/

[9] A. Dijk, E. Halisdemir, C. Melella, A. Schu, M. Pihelgas, and R. Meier, "LSPR23: A novel IDS dataset from the largest live-fire cybersecurity exercise," *J. Inf. Secur. Appl.*, vol. 85, Sep. 2024.

[10] S. S. Sengar, A. B. Hasan, S. Kumar, and F. Carroll, "Generative artificial intelligence: A systematic review and applications," *Multimed. Tools Appl.*, Aug. 2024.

[11] Z. B. Akhtar, "Unveiling the evolution of generative AI (GAI): A comprehensive and investigative analysis toward LLM models (2021–2024) and beyond," *J. Electr. Syst. Inf. Technol.*, vol. 11, Jun. 2024.

[12] E. Karlsen, X. Luo, N. Zincir-Heywood, and M. Heywood, "Benchmarking large language models for log analysis, security, and interpretation," *J. Netw. Syst. Manag.*, vol. 32, Jul. 2024.

[13] A. Bessey et al., "A few billion lines of code later: Using static analysis to find bugs in the real world," *Commun. ACM*, vol. 53, Feb. 2010.

[14] J. Newsome and D. X. Song, "Dynamic taint analysis for automatic detection, analysis, and signature generation of exploits on commodity software," in Proc. *Netw. Distrib. Syst. Secur. Symp. (NDSS)*, 2005.

[15] Z. Sheng, F. Wu, X. Zuo, C. Li, Y. Qiao, and L. Hang, "LProtector: An LLM-driven vulnerability detection system," Nov. 14, 2024, *arXiv:2411.06493*.

[16] M. Siavvas, I. Kalouptsoglou, E. Gelenbe, D. Kehagias, and D. Tzovaras, "Transforming the field of vulnerability prediction: Are large language models the key?," in *Proc. 32nd Int. Conf. Modeling, Anal. Simul. Comput. Telecommun. Syst. (MASCOTS)*, Krakow, Poland, 2024, pp. 1–6, doi: 10.1109/MASCOTS64422.2024.10786575.

[17] J. Haurogné, N. Basheer, and S. Islam, "Vulnerability detection using BERT based LLM model with transparency obligation practice towards trustworthy AI," *Mach. Learn. Appl.*, vol. 18, Dec. 2024.

[18] H. Pearce, B. Ahmad, B. Tan, B. Dolan-Gavitt, and R. Karri, "Asleep at the keyboard? Assessing the security of GitHub Copilot's code contributions," presented at the 2022 *IEEE Symp. Secur. Priv. (SP)*, May 2022.

[19] C. Asuai and G. Nwalozie, "Investigating and addressing security policy misconfigurations," *IOSR J. Eng.*, vol. 14, Apr. 2024.

[20] C. Benzaïd and T. Taleb, "AI for beyond 5G networks: A cyber-security defense or offense enabler?," *IEEE Netw.*, vol. 34, Nov. 2020.

[21] U. Mandal, S. Shukla, A. Rastogi, S. Bhattacharya, and D. Mukhopadhyay, "µLAM: A LLM-powered assistant for real-time micro-architectural attack detection and mitigation," *Cryptol. ePrint Arch*., Paper 2024/1978, 2024. [Online]. Available: https://eprint.iacr.org/2024/1978

[22] A. Dijk, "Detection of advanced persistent threats using artificial intelligence for deep packet inspection," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2021.

[23] M. A. Ferrag, F. Alwahedi, A. Battah, B. Cherif, A. Mechri, and N. Tihanyi, "Generative AI and large language models for cyber security: All insights you need," May 21, 2024, *arXiv:2405.12750*.

[24] C.-N. Hang, P.-D. Yu, R. Morabito, and C.-W. Tan, "Large language models meet next-generation networking technologies: A review," *Future Internet*, vol. 16, Oct. 2024.

[25] K. Stein, A. A. Mahyari, G. F. III, and E. El-Sheikh, "Towards novel malicious packet recognition: A few-shot learning approach," Sep. 17, 2024, *arXiv:2409.11254*.

[26] Z. Wu, H. Zhang, P. Wang, and Z. Sun, "RTIDS: A robust transformer-based approach for intrusion detection system," *IEEE Access*, vol. 10, 2022.

[27] Z. A. Khan, D. Shin, D. Bianculli, and L. Briand, "Guidelines for assessing the accuracy of log message template identification techniques," in *Proc. 44th Int. Conf. Softw. Eng. (ICSE)*, Jul. 2022.

[28] Z. Jiang et al., "LILAC: Log parsing using LLMs with adaptive parsing cache," *Proc. ACM Softw. Eng.*, vol. 1, Jul. 2024.

[29] Z. Ma, A. R. Chen, D. J. Kim, T.-H. P. Chen, and S. Wang, "LLMParser: An exploratory study on using large language models for log parsing," presented at the 2024 *IEEE/ACM 46th Int. Conf. Softw. Eng. (ICSE)*, Apr. 2024.

[30] J. Xu, R. Yang, Y. Huo, C. Zhang, and P. He, "DivLog: Log parsing with prompt enhanced in-context learning," in *Proc. IEEE/ACM 46th Int. Conf. Softw. Eng. (ICSE)*, Apr. 2024.

[31] V.-H. Le and H. Zhang, "Log parsing with prompt-based few-shot learning," in *Proc. 45th Int. Conf. Softw. Eng. (ICSE)*, Jul. 2023.

[32] J. Huang, Z. Jiang, Z. Chen, and M. R. Lyu, "LUNAR: Unsupervised LLM-based log parsing," Aug. 2024, *arXiv:2406.07174*.

[33] R. Vaarandi and H. Bahsi, "Using large language models for template detection from security event logs," *Int. J. Inf. Security*, vol. 24, Mar. 2025.

[34] W. Zhou et al., "Star: A system for ticket analysis and resolution," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Min*., 2017.

[35] E. Aghaei, X. Niu, W. Shadid, and E. Al-Shaer, "SecureBERT: A domain-specific language model for cybersecurity," Oct. 20, 2022, *arXiv:2204.02685*.

[36] N. Arici, L. Putelli, L. Sigalini, I. Serina, and others, "LLM-based approaches for automatic ticket assignment: A real-world Italian application," in *CEUR Workshop Proc*., vol. 3551, Nov. 2023.

[37] F. Y. Loumachi and M. C. Ghanem, "Advancing cyber incident timeline analysis through rule-based AI and large language models," Sep. 2024, *arXiv:2409.0257*2.

[38] P. Balasubramanian, J. Seby, and P. Kostakos, "CYGENT: A cybersecurity conversational agent with log summarization powered by GPT-3," Mar. 25, 2024, *arXiv:2403.17160*.

[39] F. Li, H. Lang, J. Zhang, J. Shen, and X. Wang, "PreConfig: A pretrained model for automating network configuration," Mar. 14, 2024, *arXiv:2403.09369*.

[40] O. G. Lira, O. M. Caicedo, and N. L. S. da Fonseca, "Large language models for zero touch network configuration management," Aug. 23, 2024, *arXiv:2408.13298*.

[41] Y. Mikami, A. Melnik, J. Miura, and V. Hautamäki, "Natural language as policies: Reasoning for coordinate-level embodied control with LLMs," Apr. 6, 2024, *arXiv:2403.13801*.

[42] K. Dzeparoska, J. Lin, A. Tizghadam, and A. Leon-Garcia, "LLM-based policy generation for intent-based management of applications," in *Proc. 19th Int. Conf. Netw. Serv. Manag. (CNSM)*, Oct. 2023.

[43] S. Hays and J. White, "Employing LLMs for incident response planning and review," Mar. 2, 2024, *arXiv:2403.01271*.

[44] R. Kaur, T. Klobučar, and D. Gabrijelčič, "Harnessing the power of language models in cybersecurity: A comprehensive review," *Int. J. Inf. Manag. Data Insights*, vol. 5, Jun. 2025.

[45] OWASP, "OWASP Top 10 for LLM Applications 2025," *OWASP Top 10 for LLM & Generative AI Security*. Accessed: Jan. 4, 2025. [Online]. Available: https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

[46] MISP Project, "MISP." Accessed: Jan. 4, 2025. [Online]. Available: https://www.misp-project.org/

[47] P. Rafiey and A. Namadchian, "Using LLMs as AI agents to identify false positive alerts in security operation center," *Research Square*, Nov. 2024, doi: 10.21203/rs.3.rs-5420741/v1.

[48] M. Kaheh, D. K. Kholgh, and P. Kostakos, "Cyber Sentinel: Exploring conversational agents in streamlining security tasks with GPT-4," Sep. 28, 2023, *arXiv:2309.16422*.

[49] T. Ali and P. Kostakos, "HuntGPT: Integrating machine learning-based anomaly detection and explainable AI with large language models (LLMs)," Sep. 27, 2023, *arXiv:2309.16021*.

[50] O. Oniagbi, "Evaluation of LLM agents for the SOC Tier 1 analyst triage process," M.S. thesis, University of Turku, 2024.

[51] A. Baig, "Accessing the role of artificial intelligence in information security risk management," M.S. thesis, University of Jyväskylä, 2024.

[52] L. Wang et al., "From sands to mansions: Enabling automatic full-life-cycle cyberattack construction with LLM," Jul. 24, 2024, *arXiv:2407.16928*.

[53] E. Pleshakova, A. Osipov, S. Gataullin, T. Gataullin, and A. Vasilakos, "Next gen cybersecurity paradigm towards artificial general intelligence: Russian market challenges and future global technological trends," *J. Comput. Virol. Hacking Tech.*, vol. 20, Sep. 2024.

[54] A. Dijk, R. Meier, C. Melella, and M. Pihelgas, "Locked Shields Partners Run 24 (LSPR24): A Next-Generation Cybersecurity Dataset for Blue Team Automation," *Zenodo*, Mar. 1, 2025, doi: 10.5281/zenodo.14900873.

[55] M. Cadrouil, "LLM agents in cybersecurity: A double-edged sword," I-TRACING. Accessed: Jan. 4, 2025. [Online]. Available: https://i-tracing.com/blog/llm-agents-cybersecurity/

CyCon 2025: The Next Step
17th International Conference on Cyber Conflict
C. Kwan, N. Gratzer, K. Podiņš, M. Tolppa (Eds.)
2025 © NATO CCDCOE Publications, Tallinn

Permission to make digital or hard copies of this publication for internal use within NATO and for personal or educational use when for non-profit or non-commercial purposes is granted provided that copies bear this notice and a full citation on the first page. Any other reproduction or transmission requires prior written permission by NATO CCDCOE.

Cyber Defense Through Strategic Dynamic Deception

Silvio Russo
Department of Computer Science and Engineering
University of Bologna
Bologna, Italy
silvio.russo3@unibo.it

Claudio Zanasi
Department of Computer Science and Engineering
University of Bologna
Bologna, Italy
claudio.zanasi4@unibo.it

Michele Colajanni
Department of Computer Science and Engineering
University of Bologna
Bologna, Italy
michele.colajanni@unibo.it

Abstract: In an interconnected digital world being enriched by smart devices, any passive solution for protecting infrastructure is doomed to fail. No matter how many defenses are implemented, attackers can infiltrate networked systems by exploiting technological or human vulnerabilities. In a scenario where the attackers have all the advantages, deception is a strategy that can slow down and divert attackers from penetrating the real infrastructure. Current platforms do create decoy environments to detect and divert threats, but attackers have developed methods to bypass these static deception systems. We propose a novel approach that is based on *strategic dynamic deception* where the system deceptor continuously analyzes the architecture and the traffic, and deploys credible decoy components. It leverages a combination of technologies such as virtualization, infrastructure as code, and generative AI to implement different types of decoys, such as similar system components, users, data, and network segments. The generation of small decoys should resemble the slow growth of a credible "ivy," so that it can attract even attackers who are already circulating in the system. When cyber threats are trapped in the fake portions of the infrastructure, many countermeasures can be activated, although these are outside the scope of this paper. Here we focus on strategies and technologies that can generate and deploy dynamic deception infrastructures. Our solution paves the way toward new approaches to cybersecurity that are based on proactive strategic deception.

Keywords: *deception, virtualization, generative AI, infrastructure as code*

1. INTRODUCTION

In an increasingly connected and digitized world, protecting personal and business data has become one of the most pressing challenges of our time. Cyberattacks have seen exponential growth in recent years, with a significant increase in frequency and complexity. Between 2020 and 2023, the number of attacks grew by 38% year-on-year, with 2023 alone witnessing over 493.3 million ransomware attacks, a 37% increase compared to the previous year [1]. These attacks are not only rising in quantity but are also evolving in sophistication, targeting critical infrastructure and business networks on a global scale [2]. Static, passive solutions, regardless of their efficacy, have shown inherent limitations in countering sophisticated, evolving threats [3]. Cyber deception has emerged as a pivotal and innovative strategy to bolster cyber defenses. By crafting deceptive environments designed to mislead and slow attackers, these systems augment traditional security measures, making infiltration efforts more challenging. However, the effectiveness of static deception platforms relying on predefined decoys and reactive countermeasures is diminishing as attackers refine their tactics to detect and circumvent such setups. This requires an evolution in deception strategies that can dynamically adapt to attacker behaviors.

The increasing sophistication of cyber threats has driven significant research into proactive defense strategies, with cyber deception [3] emerging as a powerful approach. Deception shifts the balance in cybersecurity by creating false realities and misleading attackers, enabling defenders to disrupt adversaries and gather intelligence. However, despite the potential of this strategy, many existing deception platforms rely on static methodologies, where decoys and deceptive elements remain largely unchanged over time. While this approach can initially deter attackers, it is ultimately limited in its effectiveness. Attackers often adapt to these predictable patterns, recognizing the decoys and bypassing them with increased sophistication. Once an attacker identifies that they are interacting with a decoy system, they may alter their strategy, enabling them to evade detection and avoid engaging with the deception system. The recognition and bypassing of static decoys undermine the long-term effectiveness of traditional deception techniques. This paper introduces a novel approach to cybersecurity based on *dynamic deception* [4]. It refers to a proactive and evolving strategy aimed at protecting systems from malicious attacks by introducing continuously changing deceptive elements. Traditional deception mechanisms, such as static honeypots, involve fixed decoy systems that remain unchanged over time. These decoys serve as traps, but due to their predictable nature, they can eventually be recognized and bypassed by experienced attackers. In contrast, dynamic deception continuously adapts its tactics by presenting varied and continuously changing decoys that closely mimic legitimate system components. This adaptive approach not only confounds attackers by blurring the line between real and fake assets but also provides more timely and insightful data on intrusion attempts, enhancing overall network security.

This approach involves the deployment of highly realistic and adaptable decoy resources such as fake data, systems, or network components strategically distributed across the network. These decoys are designed to closely mimic the behavior, appearance, and functionality of real assets, making it challenging for attackers to distinguish between genuine and counterfeit targets. By maintaining an evolving, unpredictable environment, dynamic deception increases the likelihood that attackers will engage with decoys instead of authentic resources, thereby diverting their efforts and reducing the risk of damage to critical infrastructure. Our solution is based on continuously generating and deploying decoys that closely mimic authentic system components, providing a dynamic and flexible defense layer. By utilizing technologies such as infrastructure as code (IaC) and generative artificial intelligence (AI), we are able to automate and scale the creation and deployment of decoys in a way that seamlessly integrates into existing systems.

A key component of this solution is the integration of large language models (LLMs), which are employed to generate highly realistic and contextually relevant decoys. LLMs can analyze real-time attack patterns and automatically generate new decoy resources, ensuring that attackers encounter unexpected, realistic environments that are difficult to distinguish from genuine system components. These decoys can simulate the behavior, appearance, and functionality of authentic assets with remarkable accuracy, reducing the risk of attackers bypassing them.

The rest of the paper is organized as follows: Section 2 discusses related works, Section 3 describes the methodology and the proposed solution, Section 4 presents details of the prototype and experimental tests, and Section 5 summarizes the main conclusions and outlines future work.

2. RELATED WORK

Urias et al. [5] explore cyber deception as a proactive mechanism to mitigate advanced cyber threats, including advanced persistent threats (APTs). Traditional approaches, such as static honeypots or signature-based detection, are insufficient against attackers capable of exploiting human and technological vulnerabilities. Urias et al. [5] emphasize the potential of deception to misdirect, delay, and neutralize attackers through techniques that manipulate their perception of the environment.

Their analysis categorizes deception strategies into concealment and simulation. Concealment focuses on obscuring critical assets, while simulation creates decoy environments that mimic real systems. However, these approaches often lack adaptability and fail to address scenarios where attackers can recognize and bypass static deception mechanisms. This limitation underscores the need for dynamic strategies, which are at the core of our proposed solution.

Li et al. [6] propose a framework that employs deep reinforcement learning to dynamically deploy deception strategies in container-based cloud environments. Their framework models potential attack paths using a system risk graph, which evaluates the likelihood of exploitation based on metrics such as exploit difficulty and exploit code maturity. The novelty lies in an adaptive decoy placement strategy that misguides attackers by dynamically altering perceived attack paths. Although highly effective for containerized systems, the framework's reliance on complex orchestration makes management of the deception infrastructure very difficult. Our goal is instead to implement a lightweight, incremental decoy generation system that requires minimal human effort. Such a system would be more adaptable and scalable.

Ivanova et al. [7] examine the use of deception solutions within industrial control systems. Their findings highlight the importance of deployment context, as on-site honeypots attract more targeted attacks than cloud-based ones. While effective for specific environments, such static honeypot solutions struggle to adapt to evolving threats and attacker tactics. While these systems exemplify the value of runtime adaptability, they often focus on malware-specific scenarios and predefined deception playbooks. In contrast, our approach aims to extend this adaptability by leveraging IaC and generative AI to autonomously grow decoys incrementally, mimicking system development. This allows us to trap sophisticated attackers over time, making our solution more versatile and harder to detect.

Several deception platforms [8], [9] largely focus on creating convincing decoy environments or deploying static countermeasures when threats are detected. Our work builds on these foundations by introducing a strategic growth model for deception. Our system incrementally introduces decoys that blend seamlessly into the existing infrastructure.

By leveraging modern virtualization techniques, IaC, and generative AI, our proposal makes it easier to deploy dynamic and proactive deception infrastructure, paving the way for more resilient cybersecurity strategies.

Speaking of generative AI, it is emerging as an important ally in many sectors, including healthcare [10], [11], education [12], and cybersecurity [13]. Its ability to create novel content, generate human-like text, and produce images has made it a powerful tool for businesses and researchers alike. This versatility highlights the transformative potential of generative AI, which not only boosts productivity and efficiency but also opens new possibilities for innovation and problem-solving.

In the generative AI family, LLMs are a transformative technology, utilizing deep learning to predict, generate, and understand text across diverse applications. Trained on vast datasets, they extend their impact beyond conventional natural language processing (NLP) tasks, finding applications in software development [14] and infrastructure management [15].

Despite the limited literature on the subject, recently those technologies have started being used to generate IaC [16] to help manage infrastructure [17], an innovative solution to simplify the process of setting up and scaling infrastructure, which reduces the need for manual coding and mitigates errors. By leveraging LLMs, organizations can quickly generate IaC templates based on high-level descriptions, improving efficiency and reducing the risk of human error in infrastructure management.

Diaz-de-Arcaya et al. [18], investigate the use of LLMs to automate patching and troubleshooting processes within IaC projects. Their approach focuses on assisting developers by identifying, diagnosing, and resolving configuration errors, often caused by the inherent complexity of managing multiple infrastructure layers. While their solution emphasizes automating the correction of existing IaC code, highlighting the potential of LLMs to take on this task, our approach goes further by dynamically generating entire infrastructure components from high-level requirements.

Lee et al. [19] propose an LLM-driven framework designed to generate IaC in dynamic environments. IaC has become a standard approach to automating infrastructure management, but traditional static templates often fall short when addressing the complexity and variability of modern infrastructure. The authors argue that by leveraging recent advancements in LLMs, it is possible to overcome the limitations of static IaC templates and generate more dynamic, flexible solutions that can adapt to changing environments.

Similarly, with our solution, we want to leverage the potential of combining LLMs and IaC to generate deception infrastructure that can dynamically adapt to the infrastructure.

Palavalli et al. [20] explore the use of LLMs to enhance software developer productivity. Like our approach, the authors incorporate a feedback loop mechanism in their framework, where errors and warnings from the initially generated IaC code are provided as input to the LLM agent. This iterative process enables the agent to progressively refine and enhance the code, with the aim of ultimately producing a more precise and dependable IaC template.

To the best of our knowledge, there is no existing literature that specifically explores the use of LLMs to generate deceptive components through IaC. The use of LLMs to generate IaC plans for deception could represent a novel approach, enabling dynamic, context-sensitive creation of infrastructure that mimics real environments, thereby improving the effectiveness of threat detection and enhancing the resilience of systems against adversarial attacks.

3. PROPOSED SOLUTION

Effective deception in cybersecurity relies on maintaining unpredictability and adaptability, but traditional methods remain largely static and require frequent manual updates. Here we propose a solution that dynamically manages deception environments by integrating IaC with AI to automate the generation, deployment, and continuous adaptation of configurations. The key idea is to replace static deception strategies with a system capable of interpreting the current state of the infrastructure, generating context-aware deception components, and deploying updates autonomously. By using IaC, the solution ensures a standardized, consistent, and repeatable approach to managing infrastructure. An LLM further enhances the system by generating deception elements tailored to the evolving environment, reducing the need for manual oversight.

The architecture of the system consists of three main components. The core, developed using Python, coordinates the interaction between the LLM and the IaC pipeline. This script handles the initialization, progress monitoring, and the flow of data between components, ensuring smooth execution of the workflow.

The second component is the IaC pipeline, which reads and writes Terraform [21] configuration files. It is responsible for detecting the current state of the infrastructure and saving the updated configurations after they have been generated and validated. This module ensures that the input provided to the AI model is accurate and that the output can be directly used for deployment.

The third component is the LLM integration layer, which processes prompts constructed from the Terraform files. This layer sends the prompts to an AI model and parses the responses to ensure they conform to Terraform syntax and standards.

Together, these components form a cohesive system that automates deception infrastructure updates with reliability, ensuring minimal manual intervention while maximizing adaptability to the infrastructure.

To develop our solution, we leveraged Code Llama [22], a variant of the LLaMA [23] model specifically adapted for programming tasks. While rooted in the architecture of LLaMA, it has been fine-tuned with a focus on addressing the unique demands of software development. This specialized training combines the latest advancements in LLMs with an emphasis on domain-specific expertise. One of the most important features of Code Llama is its enhanced ability to understand and work with code. By training on a rich and diverse corpus of programming languages and paradigms, the model has developed a nuanced understanding of syntax, semantics, and the contextual relationships within code. This capability significantly improves its performance in tasks such as code completion, bug detection, and offering suggestions for refactoring. Moreover, Code Llama has been optimized specifically for programming by fine-tuning it using carefully curated datasets from open-source repositories, technical documentation, and real-world programming challenges.

Another notable strength is its support for multiple programming languages. The model excels not only in widely used languages such as Python, JavaScript, and Java but also in more specialized or emerging languages like Rust, Julia, and R. This multilingual capability ensures it remains relevant to a wide range of development contexts, from mainstream applications to niche technical domains.

Adaptability is another key strength of the model. Its modular design allows it to be further fine-tuned for specific tasks or highly specialized programming needs. Whether applied to general software development or technical challenges, the model's scalable architecture ensures it remains effective and efficient even when tackling computationally demanding tasks. With these capabilities, the model offers a powerful resource for developers. By integrating general LLM advancements with deep programming expertise, it has the potential to streamline workflows, improve code quality, and support a wide array of technical applications.

A. Implementation

The goal of our solution is to develop an automated module for deploying deception configurations by leveraging LLMs to generate IaC. This approach enables dynamic and iterative management of infrastructure, optimizes operational efficiency, and reduces the potential for manual errors. The solution incorporates several key functionalities, including file reading, interaction with language models, and automated management of the Terraform lifecycle.

At the core of the solution is the integration with the LLM that processes contextual inputs to provide accurate and adaptable recommendations. This integration enhances the system's ability to address complex scenarios, offering precise and relevant solutions tailored to specific requirements.

The deployment process is fully automated, encompassing key operational phases such as environment preparation, configuration validation, change planning, and update execution. This automation guarantees consistency and precision throughout the deployment lifecycle, eliminating the variability associated with manual processes. By automating these critical tasks, the solution enhances reliability and reduces the time required for infrastructure updates. To ensure the system remains aligned with the current state of the infrastructure, the implementation incorporates asynchronous methodologies for continuous enhancement. These methodologies enable periodic updates and refinements, in a feedback loop process, ensuring that configurations remain up-to-date and optimized. This ongoing improvement process reinforces the solution's ability to adapt to dynamic environments while maintaining its overall effectiveness.

By combining these features, the proposed solution offers a robust and efficient framework for managing and deploying deception configurations. The integration of language models with automated infrastructure management not only simplifies the process but also enhances its scalability, precision, and adaptability, making it a valuable contribution to the field.

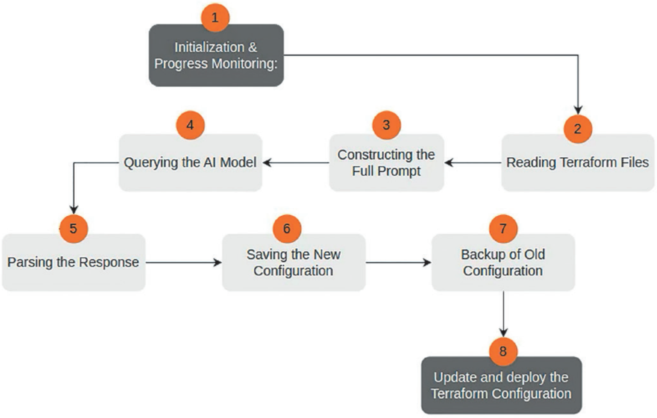
B. Execution Flow

The interaction between the various components of the system is designed to ensure a seamless flow of information and operations. The process, illustrated in Figure 1, is the following:

1. **Initialization and progress monitoring**
   The process begins with the system initializing and setting up progress monitoring. This ensures the automation workflow is tracked and managed efficiently, providing visibility into the status of each subsequent step.
2. **Reading Terraform files**
   Next, the system reads the Terraform configuration files from a specified directory. These files represent the current state of the infrastructure, and their contents are essential for building a comprehensive prompt.
3. **Constructing the full prompt**
   After reading the configuration files, the system combines the contents with any additional user-provided inputs to construct a full prompt. This unified prompt acts as the blueprint for generating a new Terraform configuration.
4. **Querying the AI model**
   The constructed prompt is then sent to the AI model, which generates an updated Terraform configuration. This step leverages the model's ability to adapt and create infrastructure definitions based on the provided context.
5. **Parsing the response**
   The AI-generated response undergoes parsing to ensure it adheres to Terraform's syntax and conventions. This step is crucial to avoid issues during deployment by verifying the correctness and usability of the generated configuration.
6. **Saving the new configuration**
   Once validated, the new Terraform configuration is saved to a file. This ensures that the generated configuration is preserved and ready for implementation.
7. **Backup of old configuration**
   Before proceeding with updates, the system creates a backup of the existing configuration. This precautionary measure safeguards against potential issues by enabling a rollback to the previous state if necessary.
8. **Update and deploy the Terraform configuration**
   Finally, the updated configuration replaces the old one and is deployed to the infrastructure using Terraform. This step updates the infrastructure to match the new desired state.

Throughout the process, error-handling mechanisms are in place. If any errors arise, such as issues in parsing, saving, or deploying, the system catches the exceptions and rolls back to a stable configuration. This approach aids in identifying and resolving problems promptly, reducing the risk of disruptions.

FIGURE 1: EXECUTION FLOW


1
Initialization & Progress Monitoring:
2
Reading Terraform Files
3
Constructing the Full Prompt
4
Querying the AI Model
5
Parsing the Response
6
Saving the New Configuration
7
Backup of Old Configuration
8
Update and deploy the Terraform Configuration

A Terraform pipeline is implemented to automate the steps required to modify and apply changes to the infrastructure. Once the system scans the specified directory to locate the relevant Terraform configuration files, ensuring that only the correct files are processed, it runs several Terraform commands in sequence: First, the *init* command initializes the working directory by setting up the necessary environment. Then, *validate* checks the configuration files for errors, ensuring they are syntactically correct and logically consistent. The *plan* command generates a preview of the changes to be made, allowing for review before execution. Finally, *apply* applies the changes to the infrastructure. By automating these steps, the pipeline reduces manual effort and minimizes the risk of human error, streamlining the management and modification of infrastructure configurations.

Finally, the system's *asynchronous automation* lies at the core of its ability to be dynamic, responsive, and highly adaptable. By decoupling tasks and executing them at different intervals, the pipeline avoids rigid scheduling constraints, ensuring flexibility and resilience in handling infrastructure updates. This asynchronous design allows the system to react to changes in real time without disrupting ongoing processes.

The periodic yet non-deterministic execution cycle ensures that the infrastructure configuration stays updated, continuously adapting to evolving conditions, making for a dynamic deception solution that evolves over time.

Each step in this process is carefully designed to ensure efficiency, reliability, and ease of use. By automating the handling of Terraform configurations and leveraging AI-generated outputs, the workflow minimizes human error and accelerates the process of managing infrastructure.

C. Model Interaction

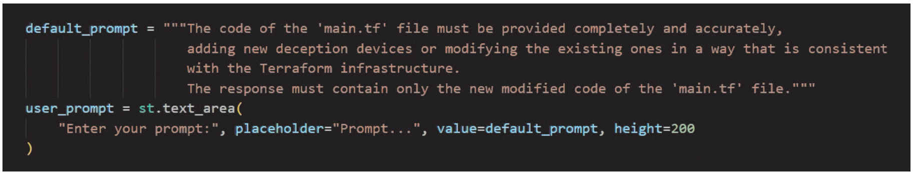
The core operation of the system relies on interaction with the LLM, which plays a critical role in dynamically generating context-aware Terraform configurations. The model's primary responsibility is to adapt the infrastructure to shifting security needs, particularly by incorporating deception elements that bolster the system's resilience against potential attacks.

To interact with the model, we leverage Ollama [24], a local AI model server that processes the prompt using advanced machine-learning techniques. Ollama's architecture is specifically designed to handle complex NLP tasks as well as context-awareness mechanisms.

Once the system collects data from the existing infrastructure, including the current state of resources, configurations, and network topologies, this information forms the basis for crafting a detailed and context-aware prompt for the AI model. The prompt is designed to communicate the specific security context of the infrastructure and the operational requirements, ensuring that the AI model understands the environment's unique needs.

The prompt, an example of which is shown in Figure 2, serves as the input for instructing the AI model to generate updated Terraform configurations. These configurations are tailored to integrate a variety of deception components, such as honeypots, decoy services, and traffic monitoring tools. Honeypots mimic real systems with intentional vulnerabilities to trap attackers, while decoy services create fake endpoints that divert malicious actors away from critical assets. By incorporating these deception elements into the infrastructure, the system ensures that it remains flexible and capable of evolving in response to new and emerging security threats.

FIGURE 2: PROMPT EXAMPLE



```
default_prompt = """The code of the 'main.tf' file must be provided completely and accurately,
                    adding new deception devices or modifying the existing ones in a way that is consistent
                    with the Terraform infrastructure.
                    The response must contain only the new modified code of the 'main.tf' file."""
user_prompt = st.text_area(
    "Enter your prompt:", placeholder="Prompt...", value=default_prompt, height=200
)
```

Additionally, there is the flexibility for the user to modify the prompt based on specific needs or preferences. This allows for a customized approach, where users can refine the prompt to address particular security concerns or infrastructure requirements.

In this way, as new threats arise or the system's operational conditions change, it continuously updates the Terraform configurations to reflect these changes, enabling the infrastructure to respond in real time to security events. This constant adaptation helps to create a resilient, self-optimizing defense system that is capable of staying ahead of adversaries and minimizing the risk of exploitation.

4. EXPERIMENTS

We conducted several experiments to evaluate our solution and the ability of the system to generate deception components. Although our solution can be easily implemented in a cloud environment using IaC, we conducted our experiments locally, using Docker containers.

For our tests, we use the 7B version of Code Llama, which provides a balanced capability to understand and generate code for problems of moderate complexity. This version offers improved accuracy compared to the smaller version, while still handling various tasks effectively without requiring excessive computational resources.

The initial infrastructure has the following components.

- Docker network: A custom Docker network (testbed network) is created to ensure that the containers can communicate with each other.
- Web server (Apache HTTP server): A basic Apache web server is deployed to serve static content. The web server listens on port 8080 externally and port 80 internally. It is configured to display a simple message that serves as the landing page.
- Database (MySQL): A MySQL database container is deployed on port 3306. This container stores application data and communicates with the other containers.
- Cache server (Redis): A Redis container is deployed to simulate a cache server. It listens on port 6379 and can be used by other components (such as the web app) for caching purposes.
- Load balancer (nginx): An nginx load balancer container is deployed to distribute incoming traffic across multiple backend servers. It listens on port 8082 and can be used to balance traffic across the web application or web server.
- Web application (Flask app): A Flask web application container is deployed and listens on port 5000. The application is used to simulate a basic web app interacting with the database and cache server.

Starting with this configuration, we test the ability of our solution to generate deception components coherent with the starting infrastructure. An example of the generated components is described below.

The 7B model generates two deception components: a honeypot that acts like a web server and a traffic logger.

The *honeypot container* is designed to simulate a vulnerable service that attackers are likely to target. Rather than being a passive decoy, the container simulates an active, vulnerable service that seems to have weaknesses that attackers may try to exploit.

The container runs a basic web server, creating a vulnerable environment that could be targeted by attackers seeking to exploit common web-based vulnerabilities such as SQL injection or command injection. This enhances the credibility of the honeypot and increases the likelihood that an attacker will engage with it.

To further enhance its believability, the honeypot is configured to log all interactions, capturing valuable information such as IP addresses, request types, and any attempted exploits. By capturing these interactions, the honeypot not only diverts attention from the real infrastructure but also provides critical intelligence that can be used to strengthen security measures. Additionally, the honeypot can simulate responses to common attack tools, allowing it to engage with attackers in a more interactive manner.

Finally, the container is configured to restart automatically, ensuring that the honeypot remains operational and continuously attracts and engages attackers. This guarantees that the honeypot will always be available as a decoy, serving as a persistent and reliable tool to gather data and distract attackers from the real infrastructure.

Alongside the honeypot, a *traffic logger container* is generated. The primary purpose of the traffic logger is to monitor network traffic and detect potential malicious activity or abnormal patterns indicative of an attack. By capturing network interactions, the logger provides additional context and visibility into the attacker's methods, thereby enhancing the detection and analysis capabilities of the overall security system.

The traffic logger container runs a simple command that simulates network traffic monitoring, creating a secondary point of engagement for any attackers that may attempt to interact with the system. This allows the traffic logger to monitor and record interactions with the infrastructure, which can help identify attack strategies, tools, and potential vulnerabilities that could be exploited.

In terms of network configuration, the traffic logger container is connected to the same custom network as the honeypot container, ensuring that all network interactions within the environment are properly captured. The combination of the honeypot and traffic logger provides a dual-layered deception approach, where one component engages attackers by acting as a fake target, and the other monitors and logs the traffic and interactions for analysis.

A. Model Comparison

While the majority of experiments were conducted using the Code Llama model due to its accessibility and documented efficacy, the modular design of our system facilitates the replacement of the model. This flexibility enables seamless integration and testing of alternative models within the system.

To assess the comparative performance of older models across various parameter scales, we conducted a series of experiments using models with parameter sizes ranging from less than 3 billion (B) to over 13B. The models tested were categorized into three main groups based on their parameter sizes.

Models with ≤ 3B Parameters

In this case, the test was conducted with Llama3.2 [25], with 3B and 5B parameters.
Models within this range demonstrated notable challenges in understanding requests. These models frequently experienced difficulties in comprehending the underlying context of a problem, which hindered their ability to generate appropriate infrastructure components.

In terms of *accuracy*, the generated code frequently exhibited a higher probability of containing syntactic or logical errors, which led to a reduction in the models' dependability when tasked with producing precise code, particularly in more intricate or specialized situations.

Moreover, these models often required targeted *fine-tuning* to improve their performance and deliver meaningful results. Without such fine-tuning, their ability to handle complex tasks was significantly limited, reducing their overall usefulness in demanding applications.

Models with 7B Parameters

In this case, we tested two different models: Code Llama and StarCoder2 [26].

The 7B parameter model, used most frequently during the tests, improved context understanding and problem-solving for tasks of medium complexity. Compared to smaller models, it showed a better *context understanding* and produced more accurate code. However, despite these enhancements, it struggled with more advanced scenarios and tended to generate repetitive components.

In terms of *accuracy*, this model produced fewer errors than its smaller counterparts. While it delivered better performance out of the box, it remained limited by its knowledge base and reasoning capabilities. Although it is more reliable for moderately complex tasks, fine-tuning could further enhance its precision, particularly in specialized domains.

In terms of performance, the model did reasonably well even without significant fine-tuning. However, there is a clear potential for improvement through targeted fine-tuning, which could increase its accuracy and enable it to handle more nuanced and demanding tasks more effectively.

Models with ≥ 13B Parameters

In this case, we again used the Code Llama and StarCoder2 models, with 13B and 15B parameters respectively.

These models showed a remarkable ability to grasp complex dependencies and thoroughly understand the context of the problem. Their deep contextual awareness enabled them to generate accurate and reliable code across a diverse range of tasks, significantly outperforming smaller models in this regard.

In addition to their strong contextual understanding, these models displayed an advanced level of knowledge, particularly when dealing with complex infrastructure provided as input. This advanced comprehension allowed them to produce high-quality code with fewer logical errors and inconsistencies, making them highly effective for demanding applications.

Another notable advantage is their impressive performance without requiring extensive fine-tuning. Despite being used in their default configurations, these models delivered exceptional results. However, this superior performance came at the cost of increased computational demands, in terms of both data consumption and processing power.

5. CONCLUSION

This paper highlights the transformative potential of LLMs in enabling dynamic deception within IaC generation, showcasing their ability to revolutionize infrastructure management workflows. By leveraging advanced models such as LLaMA and Code Llama, the project introduces a novel integration of deception techniques into Terraform configuration management. This approach represents an advance in infrastructure security, introducing an element of unpredictability that deters malicious activities while optimizing operational workflows.

The proposed solution addresses the inherent complexities and challenges associated with IaC, underlining the critical role of dynamic deception in enhancing both security and efficiency. By combining intelligent file handling, contextual processing using LLMs, and fully automated deployment mechanisms, the system generates configurations that appear plausible yet are intentionally misleading. These deceptive configurations serve as a proactive defense strategy, making unauthorized actions more difficult and less effective while safeguarding the integrity of critical infrastructure. The implementation of dynamic deception through LLMs also highlights the intersection of AI and cybersecurity, showcasing the potential of these technologies to redefine traditional security paradigms. By utilizing the advanced reasoning and adaptability of models like Code Llama, this project exemplifies how AI-driven solutions can address complex and evolving threats in digital environments. This fusion of AI and cybersecurity not only enhances the security posture of infrastructure but also establishes a foundation for more innovative and resilient approaches to managing and protecting critical systems.

As LLM technologies continue to evolve, their applications in dynamic deception are likely to expand, opening new possibilities for securing increasingly complex and distributed digital infrastructures. The ability to create adaptive, misleading configurations that align with real-world scenarios while deterring malicious actions represents a groundbreaking shift in how infrastructure is managed and protected.

Some areas require a deeper investigation. In particular, to improve the generation of deception components, one key area of focus is *training with specialized datasets*. The goal is to enhance the model by exposing it to data specifically related to cybersecurity deception techniques and Terraform code. By incorporating these domain-specific datasets, the model will gain a deeper understanding of the complexities and intricacies associated with deception tactics. This specialized knowledge will help the model generate more effective and realistic deception components that align closely with the specific needs of real-world security environments, improving the overall efficacy of cybersecurity defenses.

In general, future developments will focus on improving the *context awareness* of the model. Currently, models may generate configurations that are technically correct but lack an understanding of the specific context in which those configurations will be deployed. By improving context awareness, the model will be better equipped to understand the environment in which its generated configurations will operate. By adapting the configurations to the specific requirements of different environments, the model will generate more relevant and effective deception measures, helping organizations create dynamic and contextually appropriate security layers that are more difficult for attackers to bypass.

This paper sets the stage for future research, emphasizing the importance of leveraging the growing capabilities of LLMs to address emerging challenges in infrastructure security and optimization. Through this approach, dynamic deception can become a cornerstone of resilient and secure infrastructure management in the digital age.

ACKNOWLEDGMENTS

This work was partially supported by project SERICS (PE00000014), under the MUR National Recovery and Resilience Plan funded by the European Union—NextGenerationEU.

REFERENCES

[1] Varonis. “157 cybersecurity standards and trends [updated 2024].” 2024. [Online]. Available: https://www.varonis.com/blog/cybersecurity-statistics

[2] European Union Agency for Cybersecurity (ENISA). “ENISA threat landscape 2024.” 2024. [Online]. Available: https://www.enisa.europa.eu/topics/cyber-threats/threats-and-trends/enisa-threat-landscape

[3] M. H. Almeshekah and E. H. Spafford, “Cyber security deception,” in *Cyber Deception: Building the Scientific Foundation*, S. Jajodia, V. S. Subrahmanian, V. Swarup, and C. Wang, Eds., Cham: Springer, 2016, pp. 23–50, doi: 10.1007/978-3-319-32699-3_2.

[4] L. Zhang and V. L. L. Thing, “Three decades of deception techniques in active cyber defense—Retrospect and outlook,” *Comput. Secur.*, vol. 106, p. 102288, 2021, doi: 10.1016/j.cose.2021.102288.

[5] V. E. Urias, W. M. S. Stout, J. Luc-Watson, C. Grim, L. Liebrock, and M. Merza, “Technologies to enable cyber deception,” in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, 2017, pp. 1–6, doi: 10.1109/CCST.2017.8167793.

[6] H. Li, Y. Guo, P. Sun, Y. Wang, and S. Huo, “An optimal defensive deception framework for the container-based cloud with deep reinforcement learning,” *IET Inf. Secur.*, vol. 16, pp. 178–192, 2021.

[7] S. Ivanova and N. Moradpoor, “Fake PLC in the cloud, we thought the attackers believed that: How ICS honeypot deception gets impacted by cloud deployments?” in *Proc. IEEE Int. Conf. Factory Commun. Syst. (WFCS)*, 2023, pp. 1–4, doi: 10.1109/WFCS57264.2023.10144119.

[8] M. S. I. Sajid et al., “SODA: A system for cyber deception orchestration and automation,” in *Proc. Annu. Comput. Secur. Appl. Conf. (ACSAC)*, New York, NY, USA: ACM, 2021, pp. 675–689, doi: 10.1145/3485832.3485918.

[9] M. S. I. Sajid, J. Wei, E. Al-Shaer, Q. Duan, B. Abdeen, and L. Khan, “symbSODA: Configurable and verifiable orchestration automation for active malware deception,” *ACM Trans. Privacy Secur.*, vol. 26, no. 4, Nov. 2023, doi: 10.1145/3624568.

[10] Y. Shokrollahi, S. Yarmohammadtoosky, M. M. Nikahd, P. Dong, X. Li, and L. Gu, “A comprehensive review of generative AI in healthcare,” 2023, *arXiv:2310.00795*.

[11] J. Varghese and J. Chapiro, “ChatGPT: The transformative influence of generative AI on science and healthcare,” *J. Hepatol.*, vol. 80, no. 6, pp. 977–980, 2024.

[12] U. Mittal, S. Sai, V. Chamola, et al., “A comprehensive review on generative AI for education,” *IEEE Access*, 2024.

[13] A. H. Oveis, G. Meucci, F. Mancuso, F. Berizzi, and A. Cantelli-Forti, “Generative AI threats to maritime navigation using deceptive ISAR images,” *IEEE Access*, vol. 12, pp. 173800–173809, 2024, doi: 10.1109/ACCESS.2024.3500774.

[14] F. Lin, D. J. Kim, et al., “When LLM-based code generation meets the software development process,” 2024, *arXiv:2403.15852*.

[15] T. Chanus and M. Aubertin, “LLM and infrastructure as a code use case,” 2023, *arXiv:2309.01456*.

[16] Y. Huang et al., “Advancing transformer architecture in long-context large language models: A comprehensive survey,” 2024, *arXiv:2311.12351*.

[17] K. G. Srivatsa, S. Mukhopadhyay, G. Katrapati, and M. Shrivastava, “A survey of using large language models for generating infrastructure as code,” 2024, *arXiv:2404.00227*.

[18] J. Diaz-de-Arcaya, J. López-de-Armentia, G. Zárate, and A. I. Torre-Bastida, “Towards the self-healing of infrastructure as code projects using constrained LLM technologies,” in *Proc. ACM/IEEE Int. Workshop Autom. Program Repair (APR)*, New York, NY, USA: ACM, 2024, pp. 22–25, doi: 10.1145/3643788.3648014.

[19] J. Lee, S. Kang, and I.-Y. Ko, “An LLM-driven framework for dynamic infrastructure as code generation,” in *Proc. Int. Middleware Conf.*, New York, NY, USA: ACM, 2024, pp. 9–10, doi: 10.1145/3704440.3704778.

[20] M. A. Palavalli and M. Santolucito, “Using a feedback loop for LLM-based infrastructure as code generation,” 2024, *arXiv:2411.19043*.

[21] HashiCorp. “Terraform: Infrastructure as code.” 2025. [Online]. Available: https://www.terraform.io

[22] B. Rozière et al., “Code Llama: Open foundation models for code,” 2024, *arXiv:2308.12950*.

[23] H. Touvron et al., “LLaMA: Open and efficient foundation language models,” 2023, *arXiv:2302.13971*.

[24] Ollama. “Ollama: Open-source LLaMa-based chatbot framework.” 2024. [Online]. Available: https://github.com/ollama/ollama

[25] LLAMA. “LLAMA language models.” 2025. [Online]. Available: https://www.llama.com/

[26] BigCode Project. “StarCoder2: A code generation model.” 2025. [Online]. Available: https://github.com/bigcode-project/starcoder2

CyCon 2025: The Next Step
17th International Conference on Cyber Conflict
C. Kwan, N. Gratzer, K. Podiņš, M. Tolppa (Eds.)
2025 © NATO CCDCOE Publications, Tallinn

Permission to make digital or hard copies of this publication for internal use within NATO and for personal or educational use when for non-profit or non-commercial purposes is granted provided that copies bear this notice and a full citation on the first page. Any other reproduction or transmission requires prior written permission by NATO CCDCOE.

Navigating Turbulence: Understanding New GNSS Risks in Conflict Zones

Michael Felux
School of Engineering
Zurich University of Applied Sciences
Winterthur, Switzerland
michael.felux@zhaw.ch

Benoit Figuet
SkAI Data Services
Zurich University of Applied Sciences
Winterthur, Switzerland
benoit.figuet@skai-data-services.com

Vincent Lenders
Cyber-Defence Campus
armasuisse
Thun, Switzerland
vincent.lenders@armasuisse.ch

Raphael Monstein
SkAI Data Services
Zurich University of Applied Sciences
Winterthur, Switzerland
raphael.monstein@skai-data-services.com

Martin Strohmeier
Cyber-Defence Campus
armasuisse
Thun, Switzerland
martin.strohmeier@armasuisse.ch

Abstract: Aviation's dependency on global navigation satellite systems (GNSSs) has highlighted the vulnerability of aircraft to jamming and spoofing attacks, which pose significant safety and security challenges. These threats, created and exacerbated by recent geopolitical conflicts, underscore the necessity for monitoring and mitigating GNSS interference to ensure the integrity of aviation systems. This paper examines the impact of GNSS interference on civil and military aviation, focusing on its operational and safety implications. By building a new tool that exploits Automatic Dependent Surveillance-Broadcast (ADS-B) data, we identify hotspots of GNSS jamming and spoofing in conflict zones during 2024 in regions such as the Black Sea, Eastern Mediterranean, and the India–Pakistan border during peak periods. Daily detections peaked at 1,500 flights, declining to about 500 flights by late 2024.

Finally, we discuss the applicability of existing countermeasures, such as cryptographic authentication, multi-constellation GNSS usage, and data fusion with inertial systems. While promising, these methods face challenges in widespread adoption due to technical, operational, and regulatory constraints.

Our results contribute to understanding the evolving nature of GNSS threats and demonstrate the need for collaborative international efforts to develop resilient aviation navigation systems. By prioritizing GNSS interference monitoring and mitigation, the aviation sector can enhance safety, operational reliability, and preparedness against future threats.

Keywords: *GNSS, GPS, jamming, spoofing, aircraft*

1. INTRODUCTION

The insecurity of global navigation satellite systems (GNSSs), such as the Global Positioning System (GPS) and the European Galileo, has been known since at least the early 1990s [1]. A combination of a weak received signal strength and a lack of authentication (apart from the recently introduced Galileo Open Service Navigation Message Authentication, OSNMA [2]) makes it an easy target for both accidental and malicious interference, including jamming and spoofing. Despite many warnings from security researchers and increasing numbers of reported interference incidents across different applications [3], [4], [5], the global dependency on GNSSs has grown significantly during the past decade, simply due to their high utility in all sectors of life and industry.

While it has always been an open secret that GNSS interference could also impact the aviation sector, specifically aircraft navigation systems, until recently, it was restricted to a few known areas and limited to jamming, and as such was considered an occasional fact of pilot life [6]. With the Ukraine–Russia war and escalating tensions in the Middle East, this changed fast, and malfunctioning GNSS systems became a common and significant issue for both military and civil aircraft [7], [8].

A prominent example came in May 2024, when Tartu airport in Estonia was affected by interference (according to the Estonian government, the source was inside Russia), rendering its GPS-based approach procedures unavailable [9], [10]. This potentially lethal damage shut down all commercial flight services to the airport for weeks, until a GPS-free approach procedure based on ground-based distance measuring equipment (DME) could be enabled. Other incidents in civil aviation have involved spurious alerts from the ground proximity warning system, causing sudden and unsafe maneuvers [3].

In [11], the authors analyzed the evolution of GNSS jamming over one year after the onset of the war in Ukraine, which showed that the effect varied significantly regionally and over the year. Recent efforts from different researchers have focused on methods to detect the evolving threat of GNSS spoofing and its localization based on ADS-B data [12], [13], [14]. In this paper, we conduct a comprehensive analysis of GNSS interference on aircraft in various conflict zones in 2024. To do this, we exploit the information on the estimated quality and accuracy of the GNSS positions broadcast by civil and military aircraft using the Automatic Dependent Surveillance-Broadcast (ADS-B) technology to a large network of more than 7,000 ground-based receivers globally.

Our analysis shows that, at its peak, more than 1,500 flights have been affected daily near conflict regions from the Black Sea to the Eastern Mediterranean to the India–Pakistan border. Beyond the impact on the affected aircraft, this approach provides insights into the conflict areas subject to jamming and spoofing, the types of interference techniques used, and the origins of the interferences.

To summarize, this paper makes the following contributions:

- We present a new tool to detect GNSS interference, both jamming and spoofing, on a global scale. The tool exploits the ADS-B technology widely deployed in civil and military aircraft worldwide.
- We analyze one year of incidents using crowdsourced open ADS-B data. We find hotspots in war and crisis areas.
- We analyze the impact on civil aviation in these airspaces by quantifying the interference, including the approximate origin, and discussing several safety-relevant case studies.
- Finally, we discuss possible avenues forward to secure GNSS. This includes academic proposals to strengthen its security and the countermeasures Galileo has deployed against spoofing.

2. BACKGROUND

We briefly describe the use of GNSS in aviation and its known security issues. We go on to explain the concept of aircraft transponder data that can be used to detect attacks on GNSSs at a global scale. We refer to the civilian version of GPS unless noted otherwise.

A. GNSSs in Aviation

Among the major global navigation satellite constellations (GPS, Galileo, GLONASS, and BeiDou), relevant standards have only been developed for GPS and GLONASS. All major aircraft manufacturers only use avionics that rely solely on GPS, leveraging its ability to determine position, velocity, and precise time. The onboard receiver estimates its position by measuring so-called pseudoranges (i.e., ranges determined based on signal propagation time measurements with a user clock unsynchronized to GPS time) from multiple satellites with known positions to an unknown user location, whether on land, at sea, or in the air. Based on this information, a set of equations with four unknowns can be formulated: the three-dimensional coordinates of the receiver and the offset of the receiver's clock relative to GPS time. By applying numerical methods, these equations are solved to determine the receiver's position and time.

For air navigation, the integrity of GPS data is continuously monitored to ensure safety. Aircraft-based integrity monitoring can be categorized into two types: receiver autonomous integrity monitoring (RAIM), which relies solely on GNSS data, and aircraft autonomous integrity monitoring, which incorporates additional sensors such as barometric altimeters, clocks, or inertial navigation systems. In RAIM, a minimum of six satellites is required to detect and exclude a faulty satellite by testing all possible combinations of satellite subsets.

1) GNSS Security Issues

In this section, we discuss the two main threat vectors to GNSSs: jamming and spoofing.

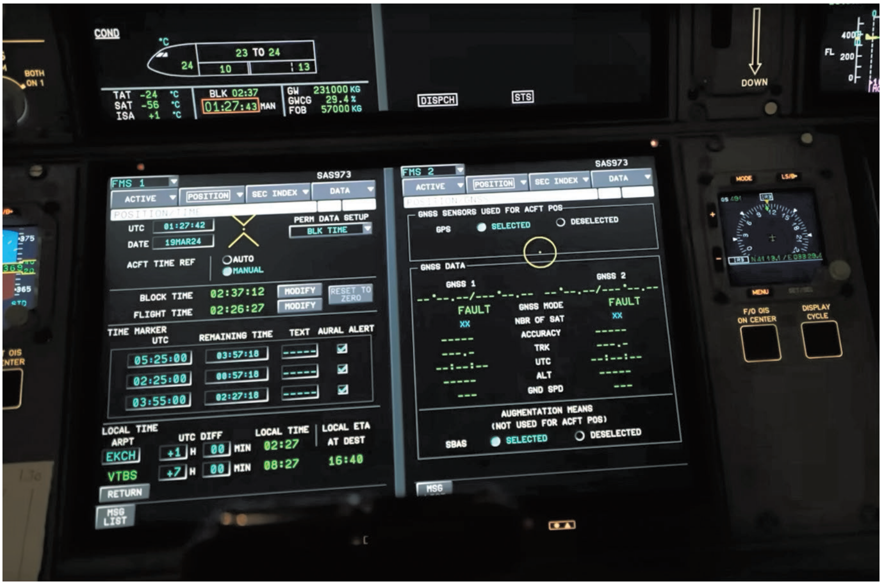
As GNSS satellite signals are extremely weak, reaching Earth below the noise floor, they can be easily interfered with even over long distances. Fundamentally, an attacker is only constrained by their power budget and line of sight to the target. Aircraft, particularly at cruising altitudes, can thus be targeted at long distances, in extremis up to the radio horizon of about 300 miles.

Jamming

Jamming involves the intentional transmission of radio frequency signals that disrupt the reception of legitimate GNSS signals. Concretely, such interference causes the receiver to lose its lock on legitimate satellites and leaves it unable to determine its position. This will typically show in the aircraft's flight instruments as "no GPS position." The position will then be estimated with less accurate means, such as conventional ground-based navigation aids like DME, VHF Omnidirectional Range

(VOR), and/or the aircraft's onboard inertial sensors, which can drift over time and can thus deviate substantially from the true position. While aircraft were designed to and are perfectly able to continue navigating without GPS, the use of precise and reliable GNSS-based navigation is a key enabler for providing airspace capacity. Given current and expected future traffic volumes, the unavailability of GNSSs can have a significant impact on capacity. Figure 1 shows the impact of jamming on an Airbus A350.

FIGURE 1: GPS JAMMING AS SEEN ON AN SAS AIRBUS A350 FROM COPENHAGEN TO BANGKOK. INTERFERENCE WAS REPORTED FROM POLAND TO THE PERSIAN GULF [15], [16]



TAT -24 °C
SAT -56 °C
ISA +1 °C
BLK 02:37
01:27:43 MAN
GW 231000 KG
GWCG 29.4 %
FOB 57000 KG

FMS 1 — SAS973
ACTIVE | POSITION | SEC INDEX | DATA
POSITION/TIME
UTC 01:27:42
DATE 19MAR24
PERM DATA SETUP: BLK TIME
ACFT TIME REF: ○ AUTO ● MANUAL
BLOCK TIME 02:37:12 MODIFY RESET TO ZERO
FLIGHT TIME 02:26:27 MODIFY

TIME MARKER UTC	REMAINING TIME	TEXT	AURAL ALERT
05:25:00	03:57:18	-----	☑
02:25:00	00:57:18	-----	☑
03:55:00	02:27:18	-----	☑

LOCAL TIME ARPT	UTC DIFF		LOCAL TIME	LOCAL ETA AT DEST
EKCH	+1 H	00 MIN	02:27	
VTBS	+7 H	00 MIN	08:27	16:40

RETURN
MSG LIST

FMS 2 — SAS973
ACTIVE | POSITION | SEC INDEX | DATA
GNSS SENSORS USED FOR ACFT POS
GPS ● SELECTED ○ DESELECTED

GNSS DATA

GNSS 1		GNSS 2
--°--.--/---°--.--		--°--.--/---°--.--
FAULT	GNSS MODE	FAULT
XX	NBR OF SAT	XX
-----	ACCURACY	-----
---.-	TRK	---.-
--:--:--	UTC	--:--:--
-----	ALT	-----
---	GND SPD	---

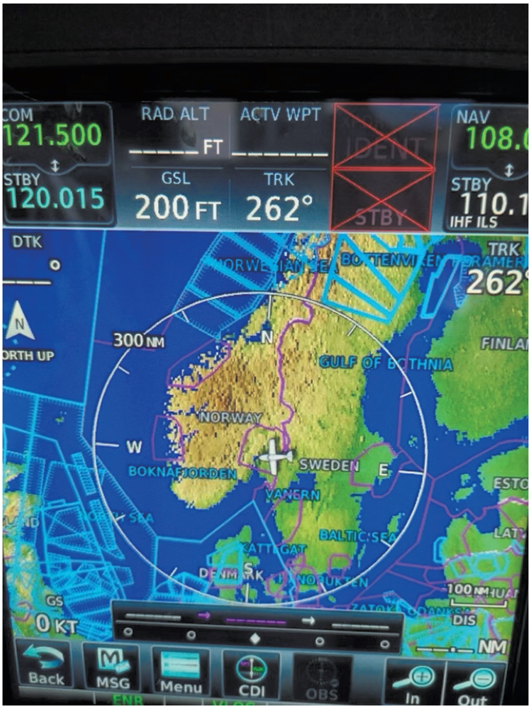
AUGMENTATION MEANS (NOT USED FOR ACFT POS)
SBAS ● SELECTED ○ DESELECTED

Spoofing

A step up in complexity from jamming, we define spoofing as manipulating or overshadowing GNSS signals to deceive receivers into calculating incorrect positions. Growing in accessibility since the early 2000s, commercial products and open-source software enable cheap, easy, and widespread execution of various spoofing techniques for many actors.

Pilots on aircraft subject to spoofing may be shown the wrong position on their flight instruments if the aircraft does not detect the GPS position error and reject it. Most passenger aircraft have proven to be effective in detecting spoofing and excluding GPS information when determining the aircraft's position. However, in less well-equipped (usually smaller) aircraft, position information may not be based on hybridization with inertial sensors and/or may be based on GPS only. An example from an avionics security laboratory experiment is shown in Figure 2.

FIGURE 2: AVIONICS SECURITY LABORATORY UNDER GPS SPOOFING. THE GARMIN SYSTEM SHOWS ITS GPS POSITION IN SCANDINAVIA DESPITE BEING LOCATED IN SWITZERLAND



B. ADS-B

The ADS-B protocol enables aircraft to broadcast their ID, position, velocity, and additional information, such as intent or urgency codes. These broadcasts occur twice per second for position and velocity updates and once every five seconds for identification. Mandated for use in U.S. and European airspace since the early 2020s, ADS-B is intended to improve the accuracy of air traffic control's (ATC) situational awareness, and it reduces system costs by replacing traditional radar systems. This protocol represents the shift toward cooperative, digital data communication networks in the next generation of ATC, enhancing and eventually replacing analog methods.

In addition to position information, parameters describing the estimated quality of the transmitted position are included in the ADS-B messages. These parameters include the Navigation Integrity Category (NIC). The Navigation Accuracy Category for Position

(NACp) indicates the maximum radius of a circle centered at the transmitted position within which the actual position is expected to be with a high probability, while the NIC specifies an integrity containment radius around the transmitted position. Both parameters are further described in the RTCA DO-260B standard document.

Most aircraft transmit ADS-B position messages based on GPS data alone, typically without hybridization or data fusion with inertial reference units, making them a particularly effective method to detect spoofing.

C. OpenSky: Crowdsourcing Air Traffic Data for Security

The OpenSky Network[1] (OSN) is a crowdsourced sensor network that collects ATC surveillance data for public access and research. Since 2013, it has continuously gathered data, offering a vast historical database of individual aircraft messages, unlike commercial flight tracking services, which aggregate them. This resource supports advancements in ATC technologies and research across various fields, from air traffic management and earth science to cybersecurity.

Starting with just eight sensors in Switzerland and Germany, the network now includes over 7,000 registered receivers worldwide. As of 2025, it has accumulated more than 12 years of data, covering over 35 trillion aircraft messages. Initially focused on ADS-B, OpenSky expanded to include further air traffic communication technologies such as Mode S, VHF, and FLARM.

The network relies on enthusiasts, academics, and institutions, with coverage limited by the line-of-sight range of individual sensors, typically 400–500 km. Growth has mirrored population density and economic development, with saturation reached in regions like Europe and the U.S. by 2024. However, significant coverage improvements continue in areas such as the Middle East, South Asia, and New Zealand, particularly at lower altitudes. OpenSky remains a vital tool for enhancing air traffic research and technologies globally, with over 650 publications depending on its data to date [17].

OpenSky provides convenient interfaces to both historical and live data for us to retrieve the necessary position and accuracy information we require for efficient detection of GNSS interference, as described in the next section.

3. IMPACT OF GNSS INTERFERENCE ON AIRCRAFT

GNSS jamming and spoofing can have severe implications for modern aircraft operations, as many aviation systems rely heavily on accurate GNSS signals for navigation, communication, time, and situational awareness. The effects and severity

[1] https://opensky-network.org

vary widely between aircraft and avionics types, and spoofing is generally more disruptive than jamming. The reason is that, whereas jamming causes the receiver to be unable to determine its position, spoofing causes the receiver to determine a faulty position. If undetected, this can cause contamination of the hybrid navigation—that is, the fused navigation obtained from multiple sources, such as inertial navigation. Below, we discuss some of the main ways in which GNSS interference can negatively impact an aircraft's systems and safety:

1. **Steering Off Course**
   The most immediate and obvious consequence of GNSS interference may be the aircraft deviating from its intended flight path if it is following a spoofed false position. This can result in the aircraft entering restricted, dangerous, or politically sensitive airspace, posing risks to operational safety and security, as has reportedly happened previously with a business jet almost entering Iranian airspace without clearance [18]. In extreme cases, this could escalate to airspace violations or military interventions. Such deviations could also lead to increased fuel consumption and delays, further compounding operational issues. It should be noted, however, that significant deviations from the intended trajectory have been very rare.

2. **Spurious Ground Proximity Warnings**
   GNSS interference can cause errors in Enhanced Ground Proximity Warning Systems (EGPWS) that rely on GNSS data to determine the aircraft's position relative to the terrain. This has reportedly already resulted in erroneous ground proximity warnings [8], [19], prompting unnecessary and potentially hazardous reactions from the flight crew. Frequent false alarms can also degrade trust in these critical safety systems, leading to delayed or complete lack of response in genuine emergencies.

3. **Reduced Navigation Performance Post-Interference**
   Even after leaving the area of direct GNSS interference, negative effects can persist as some avionics systems are seemingly unable to recover without a restart, which causes operational overheads. As a result, the aircraft may be unable to comply with the required navigation performance (RNP) standards even outside the immediately affected conflict zones, which may cause complications regarding aircraft separation and terminal navigation. We analyzed a sample of 397 transatlantic flights from Europe to America that were spoofed over the Black Sea in the summer of 2024. Of this sample, 10% were still broadcasting severely reduced navigation performance as they entered oceanic airspace.

4. **Unavailability of Approach Procedures Requiring GNSS and Precision Approaches Using GBAS/SBAS**
   GNSS-based augmentation systems, such as the Ground-Based Augmentation System (GBAS) and Satellite-Based Augmentation System (SBAS), are required for certain precision approaches. Interference can disrupt these systems and render approach procedures unavailable, as seen in incidents like the ones in Denver and Dallas in 2023.

5. **Loss of Datalink/CPDLC Capability**
   Beyond confusing the aircraft position, GNSS spoofing can lead to discrepancies between GPS time and an aircraft's internal clock. This misalignment can disrupt datalink communications, including controller–pilot data link communications (CPDLC), which require accurate time synchronization. Loss of CPDLC capability compromises efficient communication between pilots and ATC, particularly in oceanic and remote airspace, increasing workload and reliance on voice communications. This degradation in communications efficiency poses potential risks in high-traffic areas or during emergencies.

4. DETECTING GPS JAMMING AND SPOOFING USING ADS-B DATA

A. GPS Jamming

To identify GPS jamming events, two parameters from ADS-B data can be used: the NIC parameter and the NACp (similar approaches were previously discussed, for example, in [20], [21]). NIC is embedded within position messages and provides an indicator of the integrity of the navigation solution, while NACp is included in status messages and reflects the positional accuracy of the aircraft. For this analysis, NIC is preferred because it is directly embedded in position messages, which also contain the aircraft's latitude, longitude, and altitude. This makes it simpler to evaluate navigation integrity with NIC than with NACp.

In scenarios where GPS jamming is present, the NIC value typically decreases, often dropping from values above 6 (indicating reliable GPS availability) to 0 (indicating compromised navigation integrity). By monitoring NIC values, we can identify instances of GPS interference. Specifically, we aggregate the percentage of flights within a defined geographical area reporting a low NIC ($NIC < 4$) relative to those reporting a high NIC ($NIC > 6$). This ratio is calculated as follows:

$$\text{ratio} = \frac{\text{Number of flights with NIC} < 4}{\text{Number of flights with NIC} > 6 + \text{Number of flights with NIC} < 4}$$

This ratio is computed hourly to allow for temporal analysis of jamming incidents. By mapping these ratios spatially, we can identify regions affected by GPS interferences.

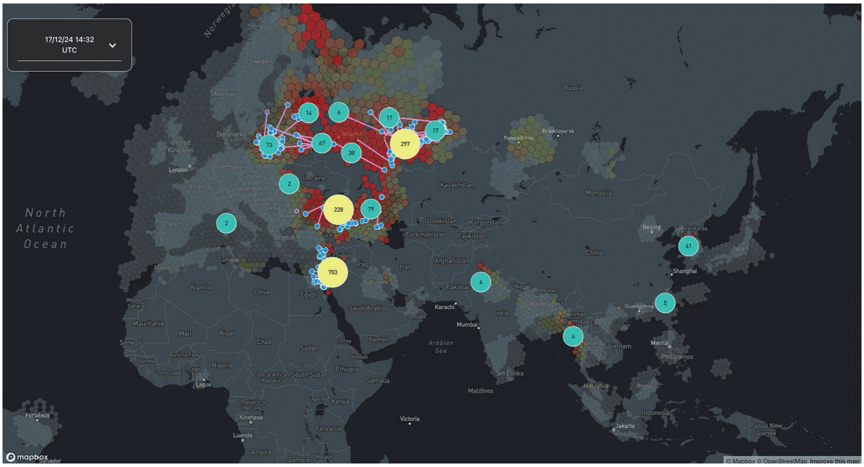
B. GPS Spoofing

We detect anomalies in the transmitted positions, such as position jumps, unrealistic speeds, or incoherent altitudes, to detect aircraft affected by spoofing. Our methodology assumes that GPS spoofing is not targeted at individual aircraft but affects multiple aircraft in the same area. To reduce noise and improve detection accuracy, we cluster the reported spoofed positions, retaining only locations reported by multiple aircraft as spoofed. This clustering approach allows us to filter out isolated anomalies and focus on genuine spoofing events that impact broader airspace regions.

C. A Tool to Monitor GPS Jamming and Spoofing Almost in Real Time

Using the techniques presented above, we have developed a website[2] that ingests ADS-B data from the OpenSky Network, identifies GPS spoofing events almost in real time, and displays GPS jamming incidents with a delay of less than two hours. Figure 3 shows the interface and display of our tool, illustrating detected jamming and spoofing interference on a single day in December 2024.

FIGURE 3: SCREENSHOT OF GPS JAMMING AND SPOOFING DETECTION MAP ON DECEMBER 17, 2024



Note: Clusters show spoofed GPS positions of aircraft, with numbers indicating the count of spoofed flights per location. Blue markers represent aircraft positions before spoofing, connected by lines to their spoofed locations. Colored hexagons indicate GPS jamming intensity, with red denoting higher levels of interference.

[2] Publicly accessible at https://spoofing.skai-data-services.com

5. EXPERIMENTAL DESIGN

A. Global Analysis

To analyze global GPS spoofing activity, we used ADS-B data from the OSN spanning January 1 to December 12, 2024. The global analysis is not geographically restricted but is limited to OSN's ground-based receiver coverage, which is more dense in regions such as Europe, North America, and parts of Asia and sparse in areas such as the polar regions and oceans. Using this dataset, we identified and characterized spoofing events using the method introduced in section 4.B.

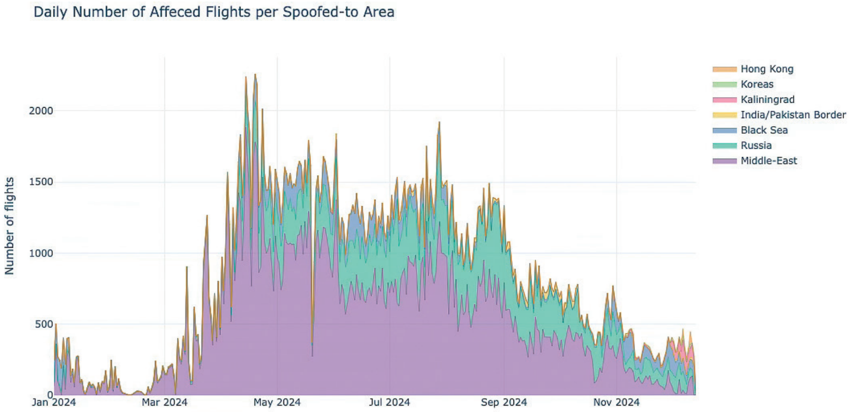
B. Kaliningrad Case Study

To analyze GPS spoofing activity around the Kaliningrad exclave, we focused on the period from December 1 to December 30, 2024, corresponding to the initial emergence of the spoofing activity in this region, which began in late November. For each identified affected flight, we also retained the last valid observation recorded before the onset of GPS spoofing.

6. QUANTITATIVE RESULTS: GNSS INTERFERENCE IN CONFLICT ZONES

While GPS jamming has been observed widely for almost two decades [22], widespread GPS spoofing affecting aviation is a relatively new phenomenon and began in earnest in the autumn of 2023. Figure 4 illustrates the rise of spoofing globally, with incidents grouped by geographical location. The daily number of detected flights plateaued around May 2024, with between 1,000 and 1,500 flights affected daily. The number decreased around September 2024 to under 500 fights affected daily. The decrease is mainly due to the decreased spoofing activities in two locations that see relatively large amounts of aircraft traffic: the Middle East and the Black Sea. Both locations have a relatively high air traffic density; thus, many aircraft are affected by it. The decrease in affected flights in the Middle East is due to the more limited operation of the spoofer. During the summer, the spoofer was active almost all day, every day. Since then, it appears that the spoofer has been active intermittently, for shorter periods of time.

FIGURE 4: NUMBER OF DETECTED UNIQUE FLIGHTS PER DAY



Note: The colors indicate the geographical region.

Individual events, such as the destruction of an oil platform in the Black Sea by the Ukrainian armed forces at the beginning of August can lead to a change in spoofing patterns. A Ukrainian spokesperson stated that the platform hosted a GNSS spoofer and was a danger to civil navigation. After the strike, the detected spoofing patterns in the area changed. While the daily number of affected aircraft decreased, it did not reduce to zero. Instead, other spoofers at different locations continued to disrupt civil air traffic.

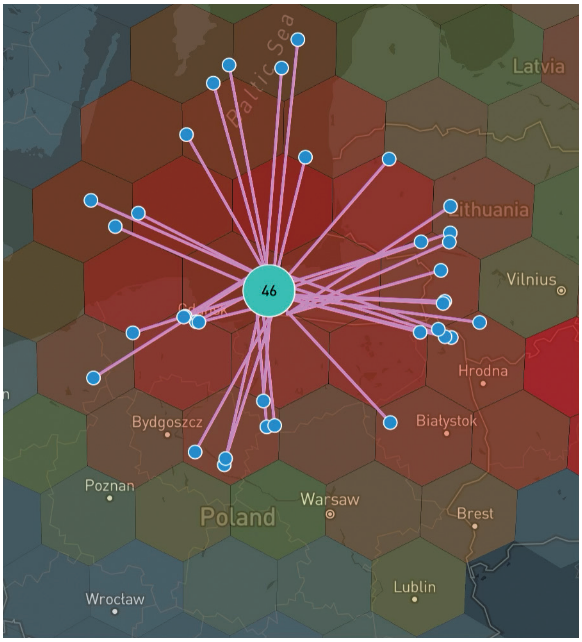
Spoofers affecting large amounts of air traffic simultaneously, such as the ones in the Middle East and Russia, are well-known, so flight crews can be briefed and prepared for a potential impact on the aircraft. However, smaller and less frequently active spoofers also pose a danger to air traffic. They might not be known to the flight crew and might take them by surprise, increasing the workload and the risk of loss of situational awareness. A noteworthy example is the GPS interference on the southern border of North Korea. On May 28–29 and June 1–2, North Korea launched hundreds of balloons filled with trash and feces toward South Korea. At the same time, GPS interference at the border was detected at scale. It started on May 28 with GPS jamming, followed by spoofing from May 30 to June 2. Even aircraft on the ground at Incheon International Airport were affected during that period. After that initial period, little interference was detected until November 5, when both jamming and spoofing activities increased again, and have continued intermittently since.

It is worth noting that GNSS interference is not limited to active conflict zones. For example, spoofing was detected on December 18 around Hong Kong and Macau. On that day, China's President Xi Jinping arrived in Macau to mark 25 years of Chinese rule over the former Portuguese enclave. This visit coincided with spoofing activities in the area on that day. In our data sample, aircraft in this area were usually not affected by spoofing, and only experienced issues on December 7 and 9.

7. CASE STUDY

Since late November 2024, we have observed significant GPS spoofing activity impacting aircraft operating over Poland, Finland, Lithuania, and Belarus. Affected aircraft reported erroneous positions localized within the Kaliningrad exclave. Figure 5 illustrates December 8, when the Live GPS Spoofing Tracker identified spoofing incidents involving 46 aircraft.

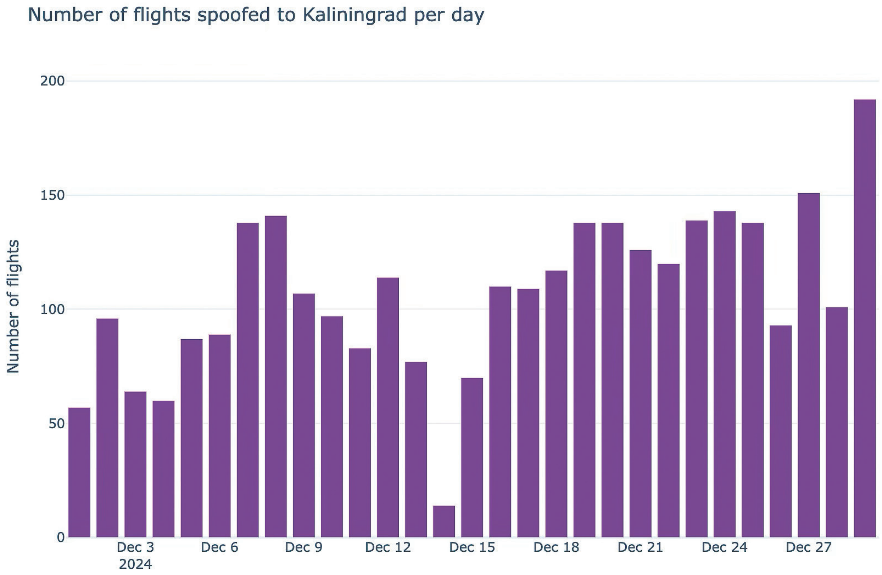
FIGURE 5: SPOOFING ACTIVITY ON DECEMBER 8, 2024, WHEN 46 AIRCRAFT WERE SPOOFED FROM KALININGRAD



Note: Blue markers represent the actual positions of the aircraft before spoofing occurred. The intensity of GPS interference is depicted by the color of the hexagons, with red indicating higher levels of detected interference.

By analyzing the historical database of the OSN, we identified over 3,000 flights affected by this spoofing activity, with as many as 150 aircraft impacted on December 30 alone. Figure 6 presents the daily count of flights detected as spoofed.

FIGURE 6: THE DAILY NUMBER OF UNIQUE FLIGHTS SPOOFED TO KALININGRAD BASED ON THE OSN HISTORICAL DATABASE

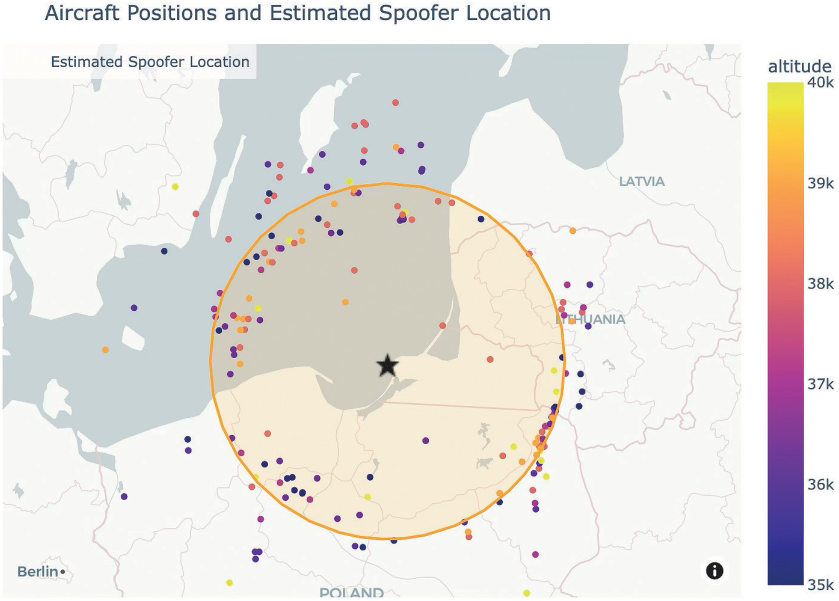


In addition to using ADS-B data to detect spoofing activity, it is also possible to approximate the location of the spoofer. By assuming a single spoofer operating from the ground and considering that radio wave propagation follows a line-of-sight principle, we infer that aircraft at the same altitude will be affected at similar distances from the spoofer.

To estimate the spoofer's position, we analyze the last valid positions of affected aircraft immediately before they were spoofed. By fitting a circle to these positions, we hypothesize that the spoofer's approximate location corresponds to the center of the fitted circle. This method provides a rough but effective estimate of the interference source.

To apply this method to our dataset, we first filter the data to include only position reports from aircraft flying at altitudes between FL350 and FL400. This altitude range is chosen to ensure that the selected aircraft are at similar flight levels, thereby maintaining consistency in their line-of-sight distance to the spoofer. Furthermore, this altitude range was selected because it contains a high density of aircraft in our dataset. Additionally, we retain only position reports where the time difference between the last valid position and the first spoofed observation is less than 5 minutes. The filtered latitude and longitude coordinates are then projected onto a Cartesian coordinate system using the EuroPP projection.

FIGURE 7: VISUALIZATION OF FILTERED AIRCRAFT DATA DEPICTING THE LAST VALID POSITIONS OF AFFECTED AIRCRAFT (MARKERS)


Aircraft Positions and Estimated Spoofer Location
Estimated Spoofer Location
LATVIA
LITHUANIA
Berlin
POLAND
altitude
40k
39k
38k
37k
36k
35k

Note: The orange fitted circle represents the estimated range of the spoofer for aircraft flying at altitudes between 35,000 and 40,000 feet, and the circle's center (star), indicates the estimated spoofer location. The markers' colors correspond to the altitude of the aircraft at the time they were spoofed.

In total, we obtained 284 data points for this analysis. The estimated spoofer location is at latitude 54.84 and longitude 19.83, in Kaliningrad, with the fitted circle having a radius of 220.4 km. This is illustrated in Figure 7, where the markers represent the last valid positions of the aircraft, and the star indicates the center of the circle, representing the likely location of the spoofer.

8. DISCUSSION

In this section, we discuss the results, their limitations, and potential countermeasures against GNSS interference in aviation.

A. The Imperative of Monitoring GNSS Interference

We argue that monitoring GNSS interference in aviation is of increasing importance, for both civil and military stakeholders. This importance is illustrated by the rapidly growing interest in up-to-date information on GNSS interference. In December, we observed almost 10,000 active users on our interference detection website, a significant amount for such a niche website that is less than a year old. At the same time, new and established organizations, such as gpsjam.org and Flightradar24,[3] have similar offerings, and the number of providers is increasing. In civil aviation, including up-to-date information in pre-flight briefings can prepare the crew for GNSS interference, reduce the surprise factor, prepare procedures to address radio frequency interference (RFI) and its associated effects, and increase situational awareness. Additionally, understanding the nature and impact of GNSS interference on aviation and all other users enables the development of technological and procedural mitigations. These advancements can make flying in the vicinity of conflict zones safer and help prevent incidents and accidents. From a military perspective, monitoring GNSS interference is essential for understanding and adapting to evolving jamming and spoofing strategies, ensuring preparedness and resilience in the face of these challenges. By prioritizing GNSS interference monitoring, aviation stakeholders can enhance safety, security, and operational effectiveness in increasingly complex environments.

B. Limitations

Detecting GNSS RFI using ADS-B data has many advantages but also several limitations.

First, it cannot detect interference in regions lacking either ADS-B coverage or active air traffic. While more and more receivers go online every day, uninhabited areas over oceans and mountain ranges are naturally difficult terrain to cover. Further, ADS-B ground receivers cluster in wealthy, industrialized regions and countries, leaving many African countries, for example, uncovered. Some of this can be mitigated using satellite-based ADS-B receivers. However, these systems cannot offer the same sensitivity and update rate [23].

Second, it should be noted that the effectiveness of this approach is contingent on aircraft flying in regions susceptible to interference, posing limitations when restricted by no-fly zones, such as those enforced by the International Civil Aviation Organization in conflict zones like Ukraine. However, open aviation data is often the

[3] https://www.flightradar24.com/data/gps-jamming

best tool available, and it is being used widely by everyone from academic researchers to air traffic controllers to mainstream data journalists.

Third, the proposed methodology cannot identify targeted spoofing attacks affecting a single aircraft. Suppose a strong military attacker focuses on a particular, high-value aircraft. In that case, they can orchestrate the signals such that only that target is affected and slowly steered off course. Here, local countermeasures on the aircraft need to be deployed to detect the attack.

Fourth, the inherent noise in ADS-B data makes it challenging to eliminate false positives and false negatives entirely. As discussed by [19], many deliberate and non-deliberate sources of interference can influence the crowdsourced air traffic data underlying our approach. While such noise makes up only a small fraction and can be filtered out in post-processing, false positives can happen. We prevent this through carefully adjusted thresholds in our algorithms and by examining interference incidents over longer periods, which increase confidence in our analysis.

C. Countermeasures

Identifying GPS interference in the cockpit of a modern airliner with its complex avionics is not always straightforward. This is because the information displayed to the pilots is often a fused position estimate drawing from various sources. Flight crews unaware of the potential of GPS interference can be caught by surprise and might need time to track the issue. This can be mitigated by providing up-to-date information to prepare the crew for such a case. Additionally, recovery of the position estimate is a crucial second step after jamming and spoofing.

We detail the state-of-the-art countermeasures grouped into approaches leveraging data fusion with other GNSS or navigation aids, cryptographic solutions, and physical and application layer solutions. The aviation sector, with its long certification and deployment timelines, faces challenges in implementing updates quickly. Modifying the Minimum Operational Performance Standards for GNSS receivers to improve (or even just begin to address) security is a significant undertaking.

1) Data Fusion

Using multiple GNSS constellations (e.g., GPS and Galileo) at the same time increases resilience against certain naive spoofing threats that only target one system. However, it is straightforward to target several systems at the same time.

Integrating GNSS data with inertial navigation systems, DME, or other ground-based systems such as VOR or non-directional beacons can provide better layers of resilience. Indeed, instead of potentially phasing out the older technologies, several

European air navigation service providers are increasing investment in them in order to reduce reliance on GPS.

2) Message Encryption and Authentication

Cryptographic approaches such as [24], [25], [26] involve encrypting and authenticating navigation messages, making it infeasible for attackers to forge GNSS signals for arbitrary spoofing. While fundamentally the best approach to securing wireless systems, standard methods are difficult to deploy in a global and open system. Military GPS does employ encryption and spread-spectrum techniques to protect against spoofing and jamming. While effective, these techniques rely on secret keys that must be securely exchanged and managed, making them infeasible for shared and global civilian systems.

The European GNSS system, Galileo, implements anti-spoofing measures based on timed efficient stream loss-tolerant authentication (TESLA) [27]. While not similarly effective as military encryption, these measures provide some authentication of the navigation data, offering enhanced protection for civilian use.

3) Physical and Application Layer Countermeasures

Techniques from the literature at the physical layer focus on detecting anomalies in radio frequency properties of the GNSS signals, such as signal strength [28] and auxiliary peaks [29], angle and direction of signal arrival [30], and validation of satellite ephemeris and timing data [29]. Using multiple antennae or receivers enhances the ability to detect interference by analyzing the spatial properties of incoming signals.

Beyond mere detection, several approaches can recover the true GNSS signal and allow the navigation system to provide the correct position. Successive interference cancellation techniques involve iterative removal of spoofed signals to isolate and recover legitimate GPS signals [31], [32]. Antenna arrays can spatially filter signals, enhancing resilience against interference and spoofing. Peripheral anti-jamming devices such as described in [16] are designed to filter out jamming signals. They are commercially available and provide protection up to certain signal strengths. However, no wireless system is entirely immune to sufficiently strong jamming attacks, particularly in contested environments or over hostile territory.

9.CONCLUSION

The escalation of GNSS interference, particularly spoofing, poses a severe threat to civil aviation safety and efficiency as increasingly severe and common recent incidents show. Urgent research is needed to enhance awareness, detection, localization, and mitigation strategies.

In this paper, we have described a novel tool that exploits ADS-B data sent by aircraft and received by crowdsourced receivers around the world to demonstrate how widespread the issue of GNSS interference is for aviation. We have identified hotspots of GNSS jamming and spoofing in several conflict zones during 2024, in particular in regions such as the Black Sea, Eastern Mediterranean, and the India–Pakistan border during peak periods. Daily detections peaked at 1,500 flights, declining to about 500 flights by late 2024.

We believe that more research is needed to better understand the potential impact of GPS interference on autopilot and autolanding systems. The impact is likely to be greatest and most direct when there are no humans in the loop. Without adequate measures, the risk of severe accidents looms large, emphasizing the imperative need for awareness, education, and proactive intervention.

REFERENCES

[1] J. R. Vasquez, “Detection of spoofing, jamming, or failure of a global positioning system (GPS),” M.S. thesis, Air Force Institute of Technology, 1992.

[2] A. Pirsiavash, A. Broumandan, and S. Kennedy, “Galileo open service navigation message authentication (OSNMA): Benefits, challenges, and limitations,” in *Proc. 37th Int. Tech. Meeting Satellite Division Inst. Navig. (ION GNSS+)*, 2024.

[3] A. Morrison, N. Sokolova, and A. Diez, “The evolving GNSS RFI threat space,” in *Proc. 36th Int. Tech. Meeting Satellite Division Inst. Navig. (ION GNSS+)*, 2023, pp. 4197–4208.

[4] J. Bhatti and T. E. Humphreys, “Hostile control of ships via false GPS signals: Demonstration and detection,” *NAVIGATION*, vol. 64, no. 1, pp. 51–66, 2017.

[5] A. Konovaltsev et al., “Interference detection and characterization with an array-based GNSS receiver using conformal antennas in maritime environments,” in *Proc. 30th Int. Tech. Meeting Satellite Division Inst. Navig. (ION GNSS+)*, 2017, pp. 2795–2811.

[6] O. Osechas, F. Fohlmeister, T. Dautermann, and M. Felux, “Impact of GNSS-band radio interference on operational avionics,” *J. Inst. Navig.*, vol. 69, no. 2, 2022.

[7] M. Felux, V. Fischer, S. Jochems, B. Figuet, and R. Monstein, “Navigating interference—Examining in-flight GNSS spoofing patterns and signal disruptions,” in *Proc. 2025 Int. Tech. Meeting Inst. Navig.*, 2025, pp. 443–452.

[8] GPS Spoofing WorkGroup, *GPS Spoofing: Final Report*, OPSGROUP, 2024.

[9] S. Jacobsen and A. Kauranen, “Estonia says Russia violates international rules with GPS interference,” *Reuters*, Apr. 30, 2024. [Online]. Available: https://www.reuters.com/world/europe/estonia-says-russia-violates-international-rules-with-gps-interference-2024-04-30/

[10] Reuters, “Finnair pauses some Estonia flights due to GPS interference,” *Reuters*, Apr. 29, 2024. [Online]. Available: https://www.reuters.com/world/europe/finnair-pauses-flights-tartu-estonia-amid-gps-interference-2024-04-29/#:~:text=%22Finnair%20will%20suspend%20its%20daily,GPS%20disruptions%20in%20the%20past

[11] M. Felux, P. Fol, B. Figuet, M. Waltert, and X. Olive, “Impacts of global navigation satellite system jamming on aviation,” *J. Inst. Navig*., vol. 71, no. 3, 2024.

[12] SkAI Data Services, “Live GPS spoofing and jamming tracker map,” 2025. Accessed: Feb. 25, 2025. [Online]. Available: https://spoofing.skai-data-services.com/

[13] Z. Liu, S. Lo, J. Blanch, and T. Walter, “GNSS spoofing detection and localization using ADS-B data,” in *Proc. 37th Int. Tech. Meeting Satellite Division Inst. Navig. (ION GNSS+)*, 2024, pp. 796–803.

[14] Kai Jansen, Matthias Schäfer, Daniel Moser, Vincent Lenders, Christina Pöpper, and Jens Schmitt, "Crowd-GPS-Sec: Leveraging crowdsurfing to detect and localize GPS spoofing attacks", IEEE Symp. on Security and Privacy (S&P), San Francisco, CA, USA, May 2018.

[15] Flightradar24, "Flightradar24 blog," Apr. 5, 2024. Accessed: Jan. 7, 2025. [Online]. Available: https://www.flightradar24.com/blog/videos/a350-longhaul-behind-the-scenes-in-the-cockpit-with-sas/

[16] Infinidome, "GPSdome2." Accessed: Jan. 7, 2025. [Online]. Available: https://infinidome.com/gps-dome-2/

[17] J. Sun, X. Olive, E. Roosenbrand, C. Parzani, and M. Strohmeier, "OpenSky report 2024: Analysis of global flight contrail formation and mitigation potential," in *Proc. Dig. Avion. Syst. Conf. (DASC)*, 2024.

[18] J. Kupietzky, "Why 20 aircraft went off course over Iraqi airspace," *Simple Flying*, Oct. 1, 2023.

[19] M. Schäfer, M. Strohmeier, M. Smith, M. Fuchs, V. Lenders, and I. Martinovic, "OpenSky report 2018: Assessing the integrity of crowdsourced Mode S and ADS-B data," in *IEEE/AIAA 37th Dig. Avion. Syst. Conf. (DASC)*, 2018.

[20] Z. Liu, S. Lo, and T. Walter, "GNSS interference detection using machine learning algorithms on ADS-B data," in *Proc. 34th Int. Tech. Meeting Satellite Division Inst. Navig. (ION GNSS+)*, 2021.

[21] B. Figuet, M. Waltert, M. Felux, and X. Olive, "GNSS jamming and its effect on air traffic in Eastern Europe," in *Proc. 10th OpenSky Netw. Symp*., 2022.

[22] D. L. Wu, C. Csar, and J. H. Salinas, "GPS jamming: A historical record from global radio occultation (RO) observations," in *37th Int. Tech. Meeting Satellite Division Inst. Navig. (ION GNSS+)*, 2024.

[23] M. Strohmeier, D. Moser, M. Schäfer, V. Lenders, and I. Martinovic, "On the applicability of satellite-based air traffic control communication for security," *IEEE Commun. Mag*., no. 9, pp. 79–85, 2019.

[24] K. Wesson, M. Rothlisberger, and T. Humphreys, "Practical cryptographic civil GPS signal authentication," *J. Inst. Navig*., vol. 59, no. 3, pp. 177–193, 2012.

[25] I. Fernández-Hernández, V. Rijmen, G. Seco-Granados, J. Simon, I. Rodríguez, and J. D. Calle, "A navigation message authentication proposal for the Galileo open service," *J. Inst. Navig*., vol. 63, no.1, pp. 85–102, 2016.

[26] S. C. Lo and P. K. Enge, "Authenticating aviation augmentation system broadcasts," in *Proc. IEEE/ION Position, Location and Navigation Symp*., 2010.

[27] A. Perrig, R. Canetti, J. D. Tygar, and D. Song, "Efficient authentication and signing of multicast streams over lossy channels," in *Proc. IEEE Symp. Secur. Priv. (S&P)*, 2000.

[28] D. M. Akos, "Who's afraid of the spoofer? GPS/GNSS spoofing detection via automatic gain control (AGC)," *J. Inst. Navig.*, vol. 59, no. 4, pp. 281–290, 2012.

[29] A. Ranganathan, H. Ólafsdóttir, and S. Capkun, "SPREE: A spoofing resistant GPS receiver," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, 2016.

[30] M. Meurer, A. Konovaltsev, M. Appel, and M. Cuntz, "Direction-of-arrival assisted sequential spoofing detection and mitigation," in *Proc. 2016 Int. Tech. Meeting Inst. Navig*., 2016.

[31] M. Eichelberger, F. Von Hagen, and R. Wattenhofer, "A spoof-proof GPS receiver," in *19th ACM/IEEE Int. Conf. Inf. Process. Sensor Netw. (IPSN)*, 2020.

[32] H. Sathaye, G. LaMountain, P. Closas, and A. Ranganathan, "SemperFi: Anti-spoofing GPS receiver for UAVs," in *Netw. Distrib. Syst. Secur (NDSS) Symp*., 2022.