Published by CCDCOE

2022

14th International Conference on Cyber Conflict: Keep Moving

T. Jančárková, G. Visky,
I. Winther (Eds.)

CYCON

IEEE
Advancing Technology for Humanity

2022
14TH INTERNATIONAL CONFERENCE ON CYBER CONFLICT: KEEP MOVING

IEEE Catalog Number: CFP2226N-PRT
ISBN (print): 978-9916-9789-0-0
ISBN (pdf): 978-9916-9789-1-7

COPYRIGHT AND REPRINT PERMISSIONS

CCDCOE Publications
Filtri tee 12, 10132 Tallinn, Estonia
Phone: +372 717 6800
Fax: +372 717 6308
E-mail: publications@ccdcoe.org
Web: www.ccdcoe.org
Layout: JDF

NATO COOPERATIVE CYBER DEFENCE CENTRE OF EXCELLENCE

The NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) is a NATO accredited cyber defence hub focusing on research, training, and exercises. Experts from most NATO nations and many partners of the Alliance across the globe work at the Centre, which is based in Tallinn, Estonia. The Centre provides a comprehensive cyber defence capability, with expertise in the areas of technology, strategy, operations, and law.

At the core of the CCDCOE is a diverse group of international experts including legal scholars, policy, and strategy experts, as well as technology researchers with military, government, and industry backgrounds.

The Centre is staffed and financed by the following NATO nations and partners of the Alliance – Austria, Belgium, Bulgaria, Canada, Croatia, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Japan, Latvia, Lithuania, Luxembourg, Montenegro, the Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, South Korea, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States.

The CCDCOE's flagships are the Tallinn Manual, the most comprehensive guide for policy advisors and legal experts on how international law applies to cyber operations; Locked Shields, the largest international live-fire cyber defence exercise; and CyCon, the annual multi-track conference bringing together key experts and decision-makers of the global cyber defence community.

The Centre is also responsible for identifying and coordinating education and training solutions in the field of cyber defence operations for all NATO bodies across the Alliance. NATO-accredited centres of excellence are not part of the NATO Command Structure.

CYCON 2022 SPONSORS

DIAMOND SPONSORS

Microsoft

FORTINET

Meta

GOLD SPONSORS

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN | iTrust

iTrust Centre for Research in Cyber Security

SILVER SPONSOR

thinklogical

A BELDEN BRAND

TECHNICAL SPONSOR

IEEE

Advancing Technology for Humanity

TABLE OF CONTENTS

INTRODUCTION

After two virtual editions, CyCon is coming back as an in-person conference. Now more than ever, we are grateful for an opportunity to meet, learn from one another, and build a community of the like-minded.

By choosing 'Keep Moving' as the central theme for CyCon 2022, the Programme Committee wanted to primarily convey, in spring 2021, the resolve not to be stopped by circumstances. Back then, our minds were primarily preoccupied with the COVID-19 pandemic. While the virus has not entirely receded, major political and security developments have since taken place in the immediate vicinity of NATO, and the theme has acquired further urgency.

The theme, of course, carries a literal meaning as well, as more and more attention is being paid to cybersecurity in the transportation industry, the maritime environment, and the supply chain, as well as to autonomous technologies.

The editors are therefore pleased to offer a collection of 23 papers chosen to best reflect all the facets of this year's theme on the three traditional CyCon tracks: law, technology, and strategy/policy.

The book opens with the topic of resilience. **Max Smeets** reviews the development and evaluates the achievements of Locked Shields, CCDCOE's flagship and the largest international cyber defence exercise. **Lee A. Bygrave** counterposes cyber resilience and cyber security as regulatory aspirations. **John Morgan Salomon**, using his rich practical experience, recommends how to make the best use of public-private partnerships.

Three papers follow with thought-provoking arguments on cyber defence and deterrence. **Martin C. Libicki** challenges deterrence-by-punishment and explores an alternative approach: obnoxious deterrence. **Erica D. Lonergan** and **Mark Montgomery** offer an assessment framework and policy recommendations to enhance deterrence and foster meaningful cooperation among NATO allies on offensive cyber operations. **Michael Bátrla** and **Jakub Harašta** focus on the value of cyber operations in disrupting the ransomware ecosystem. **Miguel Alberto Gomez** and **Gregory Winger** further develop the theme of offensive cyber operations, exploring the role of third-party countries in cyber conflict.

Several papers explore the impact of emerging and disruptive technologies on the way conflicts in cyberspace are or will be conducted, including the legal implications. **Christopher E. Whyte** studies different manifestations of AI-built intelligence in the

cyber conflict decision-making loop, while **Lennart Maschmeyer** takes a critical look at the exploitation of lethal autonomous weapons. **Arun Mohan Sukumar** argues that the growing adoption of autonomous threat-detection technologies will significantly influence state responsibility in international law, specifically by raising the duty of care demanded by the due diligence principle in cyberspace. **Tsvetelina J. van Benthem** examines the legal aspects of the unintended outcomes of using autonomous technologies in targeting under the law of armed conflict.

The discussion on the development of legal norms applicable in cyberspace continues with three more papers. In the first, **Petr Stejskal** and **Martin Faix** set out to bridge the legal gaps surrounding deceptive actions by states during cyber operations. In the following paper, **Sebastian Cymutta**, **Marten Zwanenburg**, and **Paul Oling** address legal questions pertaining to the use of biometric data during multinational military operations, with a focus on EU-led ones. Finally, **Seth W. Dilworth** and **D. Daniel Osborne** venture as far as space and address the dissonances between regulatory frameworks governing cyber operations and space assets.

Michael L. Thomas takes us to the sea domain, sounding the alarm on the vulnerability of maritime shipping to land-based cyber attacks. Following up on issues of cybersecurity in the transportation industry, **Mikko Kiviharju**, **Christina Lassfolk**, **Sanna Rikkonen**, and **Hannu Kari** investigate cybersecurity challenges faced by the railway communication systems of the future from a cryptographer perspective. **Georg Baselt**, **Martin Strohmeier**, **James Pavur**, **Vincent Lenders**, and **Ivan Martinovic** analyse satellite communication vulnerabilities in the aviation domain. **Sampath Rajapaksha**, **Harsha Kalutarage**, **M. Omar Al-Kadri**, **Garikayi Madzudzo**, and **Andrei V. Petrovski** present a context-aware intrusion detection system suitable for deployment in automobiles.

Joseph A. J. Ross, **Kimberly Tam**, **David J. Walker**, and **Kevin D. Jones** explore the use of virtualized environments for multi-objective optimization in maritime sites. **Siddhant Shrivastava**, **Francisco Furtado**, **Mark Goh**, and **Aditya Mathur** provide insight into the preparation and execution of cyber defence exercises focused on the protection of critical infrastructure and outline how the use of digital twins can assist the design of such exercises.

Emre Halisdemir, **Hacer Karacan**, **Mauno Pihelgas**, **Toomas Lepik**, and **Sungbaek Cho** then raise the issue of data obsolescence and the problem it poses for machine-learning-based intrusion detection systems; they also offer a solution that builds upon the Locked Shields exercise. In a similar vein, **Christopher Molloy**, **Philippe Charland**, **Steven H. H. Ding**, and **Benjamin C. M. Fung** introduce a phenotype-

based malware decomposition system for malware triage aimed at facilitating malware analysis and improving cyber threat intelligence.

Keir Giles and **Kim Hartmann** conclude the book with a study of NATO and allied approaches to 5G network security and supply chain challenges in the past couple of years and how these affect our preparedness for a cyber conflict.

As is the CyCon tradition and rule, all papers published in the proceedings have been subjected to a double-blind peer review by members of the CyCon Academic Review Committee. We are grateful to the reviewers, many of whom have been loyal supporters of CyCon for several years now and always found the time in their busy academic and professional schedules to help us make the final selection. We also want to thank the Institute of Electrical and Electronic Engineers (IEEE) and its Estonian section for their unwavering support and technical sponsorship of the CyCon proceedings. We likewise appreciate the effort of all the authors who responded to the 2022 call for papers, and we trust that even those who have not been selected to present at the conference will have benefited from the comments received on their research.

Naturally, this volume would not have been possible without the contribution of many other people. A heartfelt thank you goes (in alphabetical order) to Liis Poolak, Michaela Prucková, and Jaanika Rannu for organizational and moral support, and to Henrik Beckvard, Marius Gheorghevici, Davide Giovannelli, Keiko Kono, Lauri Lindström, Piret Pernik, Kārlis Podiņš, Ann Väljataga, and Jan Wünsche for their invaluable editorial assistance.

THE EDITORS

Cycon 2022 Programme Committee:

- Taťána Jančárková, chair, chief editor of the proceedings
- Cmdr. Davide Giovannelli, co-chair, law track
- Dr Keiko Kono, co-chair, law track
- Maj. Gábor Visky, co-chair, technology track
- Ingrid Winther, co-chair, strategy/policy track
- Jan Wünsche, co-chair, strategy/policy track
- Maj. Vasileios Anastopoulos
- Henrik Paludan Beckvard
- Amy Ertan
- Capt. Costel-Marius Gheorghevici
- Maj. Emre Halisdemir

- Maj. Dobrin Mahlyanov
- Piret Pernik
- Kārlis Podiņš
- Lisa Catharina Schauss
- Ann Väljataga

Academic Review Committee Members for CyCon 2022:

- Liisi Adamson, NATO CCDCOE
- Siim Alatalu, Information System Authority, Estonia
- Maj. Geert Alberghs, Ministry of Defence, Belgium
- Maj. Vasileios Anastopoulos, NATO CCDCOE
- Henrik Paludan Beckvard, NATO CCDCOE
- Jacopo Bellasio, RAND Europe, Belgium
- Dr Bernhards Blumbergs, CERT.LV, Latvia
- Prof. Thomas Chen, City, University of London, United Kingdom
- Sungbaek Cho, NATO CCDCOE
- Dr Sean Costigan, George C. Marshall Center for European Security Studies, Germany
- Sebastian Cymutta, NATO CCDCOE
- Lt. Col. Arthur Dalmijn, Ministry of Defence, Netherlands
- Samuele De Tomas Colatin, NATO CCDCOE
- Dr Thibault Debatty, Royal Military Academy, Belgium
- Dr Helen Eenmaa-Dimitrieva, University of Tartu, Estonia
- Amy Ertan, NATO CCDCOE and Royal Holloway, University of London, United Kingdom
- Cmdr. Jacob Galbreath, NATO CCDCOE
- Dr Kenneth Geers, Very Good Security, United States
- Capt. Costel-Marius Gheorghevici, NATO CCDCOE
- Keir Giles, Conflict Studies Research Centre, United Kingdom
- Cmdr. Davide Giovannelli, NATO CCDCOE
- Dr Michael Grimaila, Air Force Institute of Technology, United States
- Maj. Emre Halisdemir, NATO CCDCOE
- Dr Jonas Hallberg, Swedish Defence Research Agency, Sweden
- Dr Jakub Harašta, Masaryk University, Czech Republic
- Jason Healey, School of International and Public Affairs, Columbia University, United States
- Dr Trey Herr, Harvard University, United States
- Steven Hill, National Security Council, Washington, D.C., United States
- Prof. David Hutchison, Lancaster University, United Kingdom
- Ion Alexandru Iftimie, Cyber Security Cluster of Excellence, Romania

- Gabriel Jakobson, Altusys Corporation, United States
- Taťána Jančárková, NATO CCDCOE
- Dr Kevin Jones, Plymouth University, United Kingdom
- Kadri Kaska, NATO CCDCOE
- Dr Sokratis Katsikas, NTNU, Norway
- Dr Panagiotis Kikiras, AGT R&D GmbH, Germany
- Dr Keiko Kono, NATO CCDCOE
- Dr Csaba Krasznay, National University of Public Service, Hungary
- Lt. Col. Franz Lantenhammer, NATO CCDCOE
- Ivan Lee, Singapore University of Technology and Design, Singapore
- Dr Lauri Lindström, NATO CCDCOE
- Dr Kubo Mačák, International Committee of the Red Cross, Switzerland
- Youngjae Maeng, NATO CCDCOE
- Dr Olaf Maennel, Tallinn University of Technology, Estonia
- Maj. Dobrin Mahlyanov, NATO CCDCOE
- Dr Matti Mantere, Luminor Bank, Estonia
- Dr Paul Maxwell, Army Cyber Institute, United States
- Maj. Markus Maybaum, Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie, Germany
- Dr Stefano Mele, Italian Atlantic Committee, Italy
- Tomáš Minárik, NÚKIB, Czech Republic
- Dr Jose Nazario, Fastly, United States
- Dr Lars Nicander, Swedish National Defence College, Sweden
- Lt. Col. Gry-Mona Nordli, Norwegian Armed Forces, Norway
- Dr Alexander Norta, Tallinn University of Technology, Estonia
- Dr Kathrin Nyman-Metcalf, Tallinn University of Technology, Estonia
- Dr Sven Nõmm, Tallinn University of Technology, Estonia
- Cmdr. Rónán O'Flaherty, NATO CCDCOE
- Maj. Erwin Orye, Belgian Armed Forces, Belgium
- Dr Anna-Maria Osula, Tallinn University of Technology, Estonia
- Capt. Baris Egemen Özkan, Turkish Naval Forces, Turkey
- Dr Piroska Páll-Orosz, Ministry of Defence, Hungary
- James Pavur, University of Oxford, United Kingdom
- Piret Pernik, NATO CCDCOE
- Mauno Pihelgas, NATO CCDCOE
- Kārlis Podiņš, NATO CCDCOE
- Dr Narasimha Reddy, Texas A&M University, United States
- Lt. Col. Anastasia Roberts, NATO SHAPE, Belgium
- Olena Roraff, NATO CCDCOE
- Lt. Col. Kurt Sanger, Department of Defense, United States
- Lisa Catharina Schauss, NATO CCDCOE

- Maj. Johan Sigholm, Swedish Defence University, Sweden
- Lt. Col. Massimiliano Signoretti, Italian Air Force, Italy
- Dr Max Smeets, ETH Zurich, Switzerland
- Dr Edward Sobiesk, Army Cyber Institute, United States
- Dr Tim Stevens, King's College London, United Kingdom
- Maj. Damjan Štrucl, NATO CCDCOE
- Dr Jens Tölle, Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie, Germany
- Grete Toompere, NATO CCDCOE
- Dr Risto Vaarandi, Tallinn University of Technology, Estonia
- Ann Väljataga, NATO CCDCOE
- Lt. Col. Berend Valk, NATO CCDCOE
- Lt. Juraj Varga, NATO CCDCOE
- Dr Adrian Venables, Tallinn University of Technology, Estonia
- Maj. Gábor Visky, NATO CCDCOE
- Dr Laurin Weissinger, Tufts University, United States
- Dr Christopher Whyte, Virginia Commonwealth University, United States
- Cmdr. Michael Widmann, NATO CCDCOE
- Ingrid Winther, NATO CCDCOE
- Jan Wünsche, NATO CCDCOE
- Philippe Zotz, Luxembourg Armed Forces, Luxembourg

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

The Role of Military Cyber Exercises: A Case Study of Locked Shields

Max Smeets
Senior Researcher
Center for Security Studies
ETH Zurich, Switzerland

Abstract: What are the opportunities and limits of conducting military cyber exercises? To answer this question, I present a case study analysis of Locked Shields, the largest international cyber defense exercise in the world. Relying on non-public After Action Reports, interviews, and publicly available material, this study highlights three main challenges: the teaching of more advanced technical skills in a virtual environment; the planning of long-term campaigns below the threshold of armed attack rather than sudden cyber attacks during a two-day event; and the promotion of NATO's principles of collective defense as part of a competitive exercise. I also discuss three benefits: the ability to practice communication procedures, particularly across teams; the possibility for experimentation at the managerial level, such as assessments of team size; and the potential strategic signaling function of an exercise in a space that lacks many other opportunities for that.

Keywords: *military cyber exercises, NATO, Locked Shields, cyber conflict, learning, signaling*

1. INTRODUCTION

This article investigates the benefits and pitfalls of running military cyber exercises. In the last 20 years, numerous reports, articles, and books have been published on military exercises and wargaming.[1] These works help us understand the general tasks performed by military exercises, such as individual training, military planning, and geopolitical messaging. Several historical accounts have also shown how the computerization of wargaming has created new opportunities for gaming and analysis.[2]

Yet, military cyber exercises have received little systematic attention. Indeed, while a great deal of research has been done on how computers can be used for a wargame,[3] much less has been written on computers as the battleground of a wargame. A potential reason for this omission is the limited availability of data. Military cyber exercises are a relatively recent phenomenon, often taking place behind closed doors, with little information publicly available about these activities.[4] It is easier to study military exercises that took place several decades ago, as information tends to be less sensitive and relevant records are more likely to be accessible.[5]

This article seeks to overcome that barrier, presenting an in-depth case study analysis of Locked Shields—the largest international cyber defense exercise in the world—using non-public material, interview data, and publicly available documentation such as media reports.[6] We assess the Locked Shields exercise across three levels: individual, operational, and strategic.

1 Peter Perla, *The Art of Wargaming: A Guide for Professionals and Hobbyists* (Naval Institution Press, 1990); Martin van Creveld, *Wargames: From Gladiators to Gigabyte*s (Cambridge University Press, 2013); Beatrice Heuser, Tormod Heier, and Guillaume Lasconjarias, eds., *Military Exercises: Political Messaging and Strategic Impact*, NATO Defense College, NDC Forum Paper Series (2018); Philip Sabin, *Simulating War: Studying Conflict through Simulation Games* (Bloomsbury Academic, 2014); Diego A. Ruiz Palmer, "Between the Rhine and the Elbe: France and the Conventional Defense of Central Europe," *Comparative Strategy* 6, no. 4 (1987): 471–512.

2 James Der Derian, "The Simulation Syndrome: From War Games to Game Wars," *Social Text* 24 (1990): 187–192; Martin van Creveld, *Wargames: From Gladiators to Gigabytes* (Cambridge University Press, 2013). For a discussion on the latest generation of computerized wargames, see Andrew W. Reddie et al., "Next-Generation Wargames," *Science* 362, no. 6421 (December 21, 2018): 1362–1364.

3 There is also a related literature on the role of commercial video games. For an account of existing cyber wargames, see also Andreas Haggman, "Cyber Wargaming: Finding, Designing, and Playing Wargames for Cyber Security Education," doctoral thesis, Royal Holloway, University of London (2019), https://pure.royalholloway.ac.uk/portal/en/publications/cyber-wargaming-finding-designing-and-playing-wargames-for-cyber-security-education(5176e1b5-db99-4fe3-8289-a3d1b358fd88).html.

4 For an overview, see Robert S. Dewar, "Cybersecurity and Cyberdefense Exercises," *CSS Cyber Defense Report* (2018), https://css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/Cyber-Reports-2018-10-Cyber_Exercises.pdf.

5 For an excellent study using historical material, see Reid Pauly, "Would U.S. Leaders Push the Button? Wargames and the Sources of Nuclear Restraint," *International Security* 43, no. 2 (Fall 2018): 151–192.

6 NATO's Cooperative Cyber Defence Centre of Excellence (CCDCOE) has granted the author access to the After Action Reports (AARs) from Locked Shields. However, the CCDCOE has not granted the author permission to quote directly from the AARs. Locked Shields AARs will not be released to the public.

This study highlights three main challenges. First, the limits of the virtual environment make it difficult for personnel to teach more advanced technical skills. Second, the short time span of cyber exercises means that they tend to overemphasize the importance of sudden, highly disruptive cyber attacks and underemphasize—and thus also underprepare for—the role of ongoing cyber activity below the threshold of armed attack, which over time can have an equal strategic impact. Third, as teams of NATO allies compete against each other in Locked Shields, it is difficult to promote NATO's principles of collective defense.

Conducting an exercise like Locked Shields can also be beneficial. First, Locked Shields shows that military cyber exercises can improve communication, particularly across teams. Second, military exercises can offer significant opportunities for experimentation at the managerial level, helping to figure out the ideal team size and the value of creating multinational teams. Third, military cyber exercises can be an important signaling instrument in a space that lacks many other opportunities for that.

The remainder of this article is presented as follows. Section 2 discusses the case study design, explaining why we selected Locked Shields as a case study and what material we use to conduct this empirical inquiry. Section 3 provides a short overview of how the exercise developed over the years. The subsequent sections systematically assess Locked Shields in relation to each level of analysis. Section 4 looks at how Locked Shields can help to improve individuals' performance. Section 5 discusses how Locked Shields can help to achieve operational objectives. Section 6 considers the exercise at the strategic level, with a focus on signaling. The final section, Section 7, discusses avenues for future research.

2. CASE STUDY DESIGN

To assess the empirical validity of each proposition about the role of military cyber exercises, we conduct a case study analysis of Locked Shields. Locked Shields is an annual cyber exercise organized by the Cooperative Cyber Defence Centre of Excellence (CCDCOE).[7] It is an embedded case study design, providing an assessment of nine exercises since 2010 (there was no exercise in 2011 or in 2020).[8] The CCDCOE is a NATO-accredited cyber defense hub that fosters the cooperation of like-minded nations. Locked Shields is considered to be the most complex international exercise on cyber conflict.[9] The official goal of the exercise is to "enable cyber security experts to enhance their skills in defending national IT systems and critical infrastructure

7 Participating nations are allowed to select their training audience, which can also include personnel from government ministries, academia, or industry.

8 Roland W. Scholz and Olaf Tietje, *Embedded Case Study Methods: Integrating Quantitative and Qualitative Knowledge* (SAGE Publications, 2002).

9 CCDCOE, "Exercises" (2020), https://ccdcoe.org/exercises/.

under real-time attacks."[10] This means it is a "most-likely" case for the individual learning imperative.[11] In other words, if our analysis of Locked Shields finds weak evidence for the individual-learning imperative, then we should have serious doubts about the general efficacy of military cyber exercise to promote individual learning.

For this case study analysis, we rely on a wide range of sources. The most important source used for this evidence-based inquiry concerns the after action reports (AARs).[12] AARs provide a detailed retrospective analysis of Locked Shields. The reports discuss the exercise's annual objectives, teams' setup, participants, scenario, scoring, and technical environment. It normally also lists a set of observations and recommendations to improve Locked Shields. The AARs from 2012, 2013, and 2014 are publicly available in the online library of the CCDCOE. The CCDCOE has granted the author access to the publicly unavailable reports from later years of Locked Shields as well. As these AARs are meant not for public release but only for the participating nations and organizers of Locked Shields, it significantly reduces the public reporting bias. To gain a comprehensive picture of the empirical subject matter, we also use public reporting, such as media articles, as well as interviews conducted.[13] The author interviewed organizers of various "tracks," as well as participants from various nations.

3. HISTORICAL OVERVIEW OF LOCKED SHIELDS

This section provides a brief overview of how the Locked Shields exercise developed over time. We can roughly distinguish between three periods. The first period, 2010–2013, was the startup and experimentation phase. The second period, 2014–2017, saw the significant growth and expansion of the exercise. The final phase, which commenced in 2018, was characterized by a greater level of maturity. Figure 1 provides a general overview of the exercise's development, displaying the number of (defending) Blue Team participating teams in Locked Shields exercises over the course of each annual event. It shows that the number of Blue Teams nearly quadrupled after 2010. There was also a 10-fold increase in the maximum number of persons on a Blue Team.

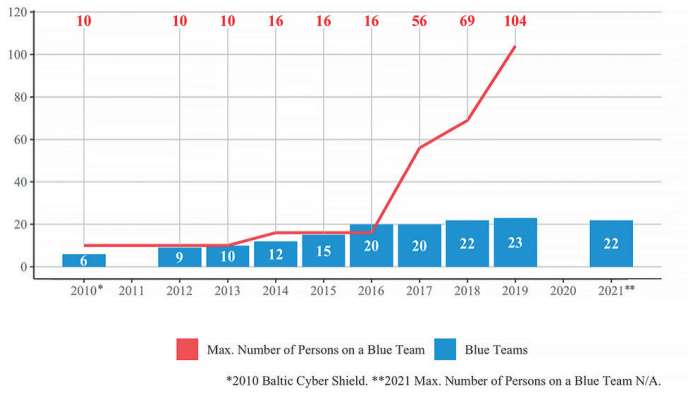
10 Ibid.

11 Harry Eckstein, "Case Study and Theory in Political Science," in *Handbook of Political Science*, ed. F. Greenstein and N. Polsby (Reading, MA: Addison-Wesley, 1975).

12 All figures and tables in this article are based on the After Action Reports.

13 I directly quote eight interviewees in this study. To give the interviewees the opportunity to freely and openly discuss their opinions, I offered them the option to remain anonymous. Interviewees identified by name have filled out a consent form.

FIGURE 1: NUMBER OF PARTICIPATING BLUE TEAMS AND MAXIMUM PERSONS PER BLUE TEAM



NATO CCDCOE started organizing the cyber defense exercises under the name "Locked Shields" in 2012. In its initial years, it was a smaller, European exercise that experimented with scenario-building and pioneered investigations of the nature of conflict in the cyber domain. Locked Shields was preceded by Baltic Cyber Shield, a one-off exercise in 2010, co-organized by the CCDCOE and the Swedish National Defence College.[14] The scenario of Baltic Cyber Shield saw Blue Teams defend energy infrastructure against militant non-governmental actors threatening attack. Baltic Cyber Shield, held shortly after the center's founding, was said to be motivated by the 2007 cyber attacks on Estonia and intended to improve states' ability to defend critical infrastructure. As the name suggests, Baltic Cyber Shield involved the Baltic States, as well as Sweden. The exercise was a technical "live-fire" exercise with a Red Team attacking six Blue Teams. Blue Teams represented rapid reaction teams (RRTs), whose task was to maintain the availability, confidentiality, and integrity of services in a small pre-built network under cyber attack. The organizational and monitoring aspects of the exercise, such as the network infrastructure, communications and scoring, and situational awareness, were the responsibilities of a Green, a White, and a Yellow Team, respectively. Approximately 100 people were involved in the 2010 exercise.

In 2012, the exercise was first held under the name of Locked Shields.[15] It involved approximately 250 participants from 10 states, expanding to include countries such as Germany, Italy, and Spain, as well as Switzerland as a one-off organizing partner. While the underlying characteristics of the exercise were adopted from the Baltic Cyber

14 CCDCOE, "Baltic Cyber Shield Cyber Defence Exercise 2010 After Action Report" (2010), https://ccdcoe.org/uploads/2018/10/BCS2010AAR.pdf.

15 CCDCOE, "Locked Shields 2012 After Action Report" (2012), https://ccdcoe.org/uploads/2018/10/LockedShields12_AAR.pdf.

Shield, the scenario now consisted of cybercriminals attacking telecommunications companies defended by nine Blue Teams. The 2013 Locked Shields welcomed new participants from the Netherlands and Poland, with a scenario focusing on a humanitarian crisis and tribal conflicts within a fictional African nation, with cyber attacks targeting aid organizations (CCDCOE 2013). It was during this third exercise that the Estonian Defense Forces joined as the exercise's technical environment host, which it remained for all subsequent exercises. The scenarios of the earlier Locked Shields exercises explored various forms of cyber attack variants and contexts, while organizational aspects were further improved.

The exercise grew significantly between 2014 and 2017.[16] Over the course of four exercises, the number of participants grew fourfold, to 800 persons and 20 Blue Teams, by 2017. France, Turkey, Hungary, and Greece joined by 2015, and in 2016, two mature global cyber powers, the United States and the United Kingdom, joined Locked Shields. In addition, more IT professionals from the private sector joined the exercise in 2014 to strengthen the Red Team.

As more nations joined, the scenario slowly developed in a more distinct direction. Starting in 2014, all Blue Teams of Locked Shields would come to represent RRTs of the fictional state "Berylia" defending against the attacking nation "Crimsonia." Berylia's research and development sector was under attack by malicious cyber actors during Locked Shields 2014 and 2015. In 2016, the exercise was expanded to include "Revalia" as a neutral nation. Blue Teams were required for the first time to practice making a legal and technical attribution of the attacks to either Crimsonia or Revalia. It was during this period of Locked Shields that inter-state cyber conflict and the political aspects of such conflict entered into clearer focus. This also led to the introduction of a strategic track in 2017.[17]

As Figure 2 shows, in 2018 and 2019, Locked Shields had over 1,000 participants, and in 2019, 23 Blue Teams were able to participate due to the larger infrastructure setup of Locked Shields. In the same year, non-NATO members Australia and Japan joined the exercise. The exercise scenario continued to focus on inter-state tensions between Berylia and Crimsonia, but now it also included a hostile Crimsonian diaspora in Berylia and NATO involvement. The Blue Teams, for the first time, played their own nations in 2018, with "forward-deployed" elements in Berylia called "Deterrent Forces" to deter Crimsonian aggression.[18] We summarize the scenario changes in Figure 3.

16 CCDCOE, "Locked Shields 2014 After Action Report" (2014), https://ccdcoe.org/uploads/2018/10/LS14_After_Action_Report_Executive_Summary.pdf; CCDCOE, "Locked Shields 2015 After Action Report" (2015), unpublished; CCDCOE, "Locked Shields 2016 After Action Report" (2016), unpublished; CCDCOE, "Locked Shields 2017 After Action Report" (2017), unpublished.

17 "Tracks" are the various activities added to the exercise. Locked Shields has technical, media, legal, and strategic tracks.

18 "Injects" consist of any information, role-play, or action delivered by the facilitation team, to one or several participants, during the exercise.

FIGURE 2: TARGET AUDIENCE LS 2012–2021 (AND BCS 2010)

*2010 Baltic Cyber Shield. **2021 "more than" 2000 participants. ***2013, 2014 Participants N/A.

FIGURE 3: TIMELINE OF LOCKED SHIELDS SCENARIO DEVELOPMENT

Scenario*	BCS 2010	LS 2012	LS 2013	LS 2014	LS 2015
Blue Team	Not specified	Telecommunication companies	International coalition of military Rapid Reaction Teams (RRTs)	Berylian RRTs	
Red Team	Not specified	Organised crime network, activist group	Boolean local extremists	Crimsonia (inconclusive attribution)	
Defended networks	Pre-built IT infrastructure of a small company	Telecommunication networks	Aid organisation systems, military systems	World Drone Expo, Berylian R&D facilities, and demonstration systems in the Expo area	Berylian drone R&D facilities, drone control systems

Scenario	LS 2016	LS 2017	LS 2018	LS 2019	LS 2021
Blue Team	Berylian RRTs		NATO deterrent forces (RRTs) stationed in Berylia representing BT nation		
Red Team	Crimsonia				
Defended networks	Berylian drone R&D facilities, drone control software	Berylian military airbase systems (etc.)	Berylian power grid and power generation, 4G public safety networks, drone control systems (etc.)	Berylian power grid and power generation, 4G public safety networks, water plant, maritime situational awareness system	Berylian air defense systems, power grid and power generation, satellite mission control system, water plant, tactical radio communication

*Blue Teams are the defenders, Red Teams are the attackers, and defended networks are the networks under attack.

4. INDIVIDUAL LEARNING

Military exercises can help to improve individuals' performance. Modern military exercises can serve at least three different functions at the individual level. First, they can train the forces, helping to reinforce old skills and generate new skills. "Train as you fight" is the often-heard motto. However, military exercises are as much about teaching actual combat skills as they are about cultivating communication, logistics, and other "softer" skills. Second, they can be used to experiment with new weapon systems and technologies. Third, they can fight off boredom and boost the morale of the soldiers.

When it comes to building up an operational capability within a military cyber command, people are the most important asset. A cyber command needs a diverse workforce, not just technical personnel—such as vulnerability analysts, developers, operators, testers, and system administrators—but also lawyers, administrators, strategists, and others.[19] At a pre-event of the CCDCOE's 2018 Conference on Cyber Conflict (CyCon) someone made the following point: for a cyber command, it is hard to find competent personnel, even harder to find competent and trained personnel, and hardest of all to find competent, trained, and experienced personnel. Certain things can only be learned on the job—which makes experience a crucial aspect.

Military cyber exercises do not perfectly match reality but can potentially help personnel gain experience and learn relevant techniques and procedures. "Techniques" refers to the behavior or actions that lead up to what the threat actor is trying to accomplish. For example, exploitation of remote services could be a technique used to move laterally within a network. Procedures are ways to leverage the technique to accomplish their objective. For example, a procedure to exploit remote services could be exploited through remote code execution (RCE) vulnerabilities—as was the case with the Mirai botnet. Military cyber exercises often take place on representations of actual networks, systems, and tools. Military cyber exercises typically make use of a range that is a collection of virtual machines hosted on the premises. Some of these ranges allow you to teach a wider range of techniques and procedures than others.

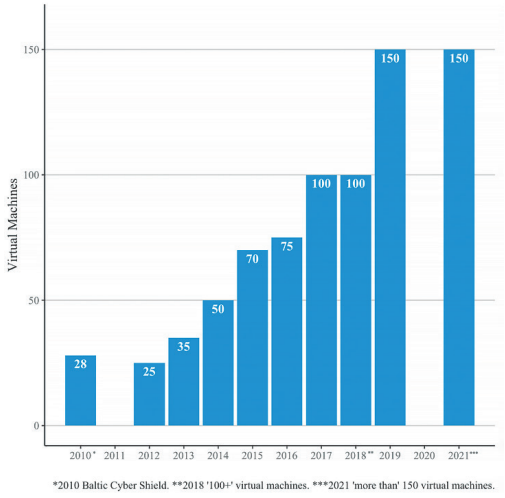
Given Locked Shields' setup and goals, the most significant learning might be expected to take place among participants in the technical track. Locked Shields managed over the years to set up a sizable technical infrastructure as a training environment. Figure 4 shows the number of virtual machines per Blue Team. The exercise grew to include over 4,000 attacks on more than 5,000 virtual systems in 2021.[20]

19 See Max Smeets, "Cyber Arms Transfer: Meaning, Limits and Implications," *Security Studies* (2022), https://www.tandfonline.com/doi/full/10.1080/09636412.2022.2041081?src=; Max Smeets, *No Shortcuts: Why States Struggle to Develop a Military Cyber-Force* (London: Hurst Publishers, 2022).

20 CCDCOE, "Locked Shields 2019 After Action Report" (2019), unpublished.

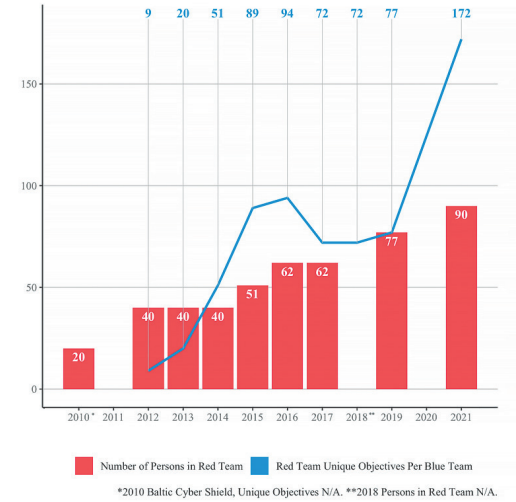
The attack campaigns also grew in complexity to increase the technical challenge for the training audience. For example, while Blue Teams had to defend against nine Red Team objectives in 2012, they had to defend against 172 in 2021. Red Team sub-teams also expanded from one to five client-side attack sub-teams to match the growing number of Blue Teams. Notably, however, the Red Teams did not grow as exponentially as the Blue or Green Teams, because having many Red Team members does not guarantee better scaling or effectiveness.

FIGURE 4: NUMBER OF VIRTUAL MACHINES PER BLUE TEAM



*2010 Baltic Cyber Shield. **2018 '100+' virtual machines. ***2021 'more than' 150 virtual machines.

FIGURE 5: RED TEAMS AND UNIQUE OBJECTIVES PER BLUE TEAM

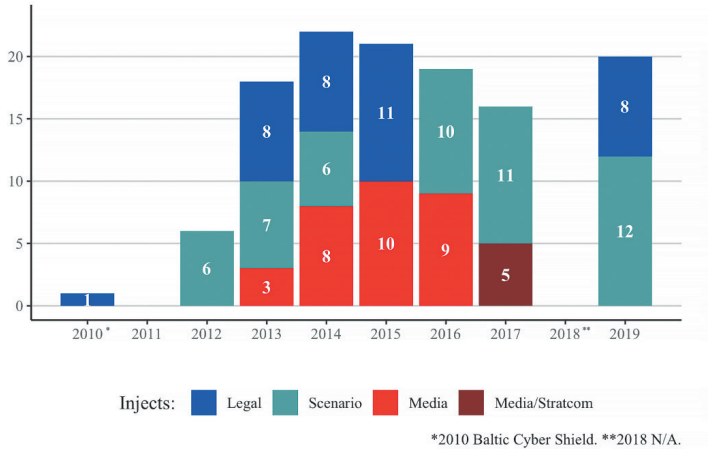


*2010 Baltic Cyber Shield, Unique Objectives N/A. **2018 Persons in Red Team N/A.

However, it turns out to be difficult to learn new technical skills at Locked Shields—especially for more experienced personnel. This is the case for a number of reasons. There is a limit to how much the virtual battleground can simulate the real environment, which means that various techniques cannot be used. For example, DDoS attacks cannot be integrated at Locked Shields because of the lack of server capacity.[21] On the offensive side, there is also little incentive to share novel techniques, as this reduces the potential effectiveness in actual scenarios. Some have argued that cyber capabilities are "single-use": once you conduct a cyber operation, exploiting a certain vulnerability and using a certain implant, it becomes known to the public and thus loses its utility.[22] The argument is that others can now defend against it. In reality, this is not the case: it often takes a long time before patches are installed and vulnerabilities closed. Implants are also frequently reused. Having said that, as successful cyber operations often rely on deception, there is great value to keeping your handiwork secret until necessary.[23] Exposure does lead to reduced efficacy. Hence, the main learning experiences came from interaction *between* the teams, rather than *within* the technical teams.

Figure 6 displays the number of "injects" over Locked Shields iterations. The figure indicates the rise in injects as the exercise grew in complexity.

FIGURE 6: NUMBER OF EXERCISE INJECTS PER BLUE TEAM



21 Interview, Rain Ottis, 2020, Zoom. The closest the infrastructure of Locked Shields came to "breaking" was in 2018 due to the configuration of the servers: Interview, Interviewee 7, 2020, Microsoft Teams.

22 Robert Axelrod and Rumen Iliev, "Timing of Cyber Conflict," *Proceedings of the National Academy of Sciences* 111, no. 4 (2014): 1298–1303.

23 Erik Gartzke and Jon Lindsay, "Weaving Tangled Webs: Offense, Defense, and Deception in Cyberspace," *Security Studies* 24, no. 2 (2015): 316–348; Max Smeets, "A Matter of Time: On the Transitory Nature of Cyberweapons," *Journal of Strategic Studies* 41 (2018): 1–2.

As Liisa Past, one of the organizers for Locked Shields for several years in various capacities, notes, "the main issue with a lot of cyber exercises—whether it is the tactical, operational, or strategic—is that you are missing other layers. You are creating a false sense of security. If the game master is writing the injects, one of the main things for me to teach wasn't about making sure techies learn technology. It was about how to learn in a team. It was about how to learn triage. And resource allocation. And learning to communicate. It is a technical exercise because it is a technical scenario. But, in the end, you are not teaching technical skills."[24] Similarly, Rain Ottis, one of the main early organizers, notes that Locked Shields is about "working together as a team under pressure. For example, how do you report to higher officials?… And what is the impact of me losing a Windows server to the rest of the group? And what to tell the media?"[25]

5. OPERATIONAL PLANNING

Second, military exercises can help to achieve operational objectives. There have been several historical cases in which military exercises and examinations directly influenced war plans. The most prominent but also most debated example is the Schlieffen Plan. Different historians have reached different conclusions about the prominence of military exercises and Moltke's *Generalstabsreise* for the direct creation of the plan.[26]

While there is considerable espionage activity in cyberspace, there is little evidence that the various militaries of the world regularly conduct cyber effect operations.[27] Yet we can assume that governments still want to plan for it operationally. Military cyber exercises can potentially help with formulating cyber-specific rules of engagement (ROE).[28] ROE are directives issued by a competent military authority that delineate the circumstances and limitations under which a nation's military forces will initiate and/or continue combat engagement with other forces encountered.[29] As Kehler, Lin, and Sulmeyer write for the United States: "[t]o the extent feasible, the [Department of Defense] has sought to apply the same principles that govern the use of kinetic weapons to the use of cyber weapons, while recognizing the special characteristics of the cyber domain and cyber weapons. This has proven to be a difficult challenge."[30] Military

24 Interview, Liisa Past, 2020, Signal.
25 Interview, Rain Ottis, 2020, Zoom.
26 Terence Zuber, *Inventing the Schlieffen Plan: German War Planning 1871–1914* (Oxford University Press, 2002); Terence Holmes, "A Reluctant March on Paris," *War in History* 2 (2001): 208–232.
27 The list of countries that conduct these operations regularly is small—no more than a dozen, based on the Council on Foreign Relations tracker.
28 Robert C. Kehler, Herbert Lin, and Michael Sulmeyer, "Rules of Engagement for Cyberspace Operations: A View from the USA," *Journal of Cybersecurity* 3, no. 1 (2017).
29 Department of Defense, Joint Publication 1–02, *Department of Defense Dictionary of Military and Associated Terms, 2010–2016*, 207, https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/dictionary.pdf.
30 Kehler, Lin, and Sulmeyer, "Rules of Engagement for Cyberspace Operations: A View from the USA."

cyber exercises can potentially help in understanding these "special characteristics" and translate this into operational planning.

Locked Shields in fact reveals how states can "experiment" with different operational setups. For example, when the cap was lifted on the maximum number of people per Blue Team, different tactics were used. Some countries experimented with having small Blue Teams one year and much larger teams the next to see how it influenced their ability to respond to the challenge effectively. As one organizer recalled, in one year the operators on one Blue Team decided that the team's best bet was to ignore any guidance from the "strategists" of the same team, even though those officers outranked them. The reasoning was that the strategists only had a limited technical understanding of the challenge. It turned out that the strategists were unable to find out whether the operators actually followed their advice. Another example of this experimentation at the operational level was in 2015, when two countries also experimented by forming one combined Blue Team.[31]

At the same time, interaction between the tracks during Locked Shields remained limited, making it harder to see the "real" effects of different operational procedures. Even in later years, advice from the strategist track was unable to "break" the exercise. As several interviewees explained, a "guardrail" has always been in place; the tracks can interact but are not fully co-dependent.[32]

Furthermore, Locked Shields, like other exercises, does not last longer than a few days and uses various injects to announce certain developments within the exercise. There is always the danger that a scenario stereotypes opponents or is based on other unrealistic behavior. The short time span of exercises seems to lead to unrealistic scenario-setting. After all, we know that unique decision-making dynamics stem from the timeframe of cyber operations.[33] First, the preparation of more advanced cyber operations takes time. Second, the duration between initial access and ultimate purpose is often a long one.[34] The installation of a backdoor in one period might lead to the dropping of malware only many months later.

Furthermore, Locked Shields revolves around a set of highly disruptive or destructive cyber attacks: as discussed, the typical scenario relates to the disruption or destruction of critical infrastructure by an adversarial actor. These types of scenarios connect to the discourse about surprise cyber attacks crippling critical infrastructure, which

31 Interview, Interviewee 6, 2020, Zoom; Interview, Interviewee 7, 2020, Microsoft Teams; Interviewee 9, 2020, Zoom; Interview, Interviewee 10, 2020, email.

32 Interview, Interviewee 5, 2020, Zoom; Interview, Interviewee 6, 2020, Zoom.

33 Matthew Monte, *Network Attacks and Exploitation* (Wiley, 2015); Max Smeets and J. D. Work, "Operational Decision-Making for Cyber Operations: In Search of a Model," *Cyber Defense Review* (Spring 2020), https://cyberdefensereview.army.mil/Portals/6/CDR%20V5N1%20-%2007_Smeets_WEB.pdf.

34 James E. McGhee, "Liberating Cyber Offense," *Strategic Studies Quarterly* 10, no. 4 (winter 2016): 46–63.

emerged in the mid-1990s and became especially prominent in the first years of the decade of the 2000s.[35] Yet today, there is a general recognition in the academic literature about the limits of cyber war as a useful concept for interpreting the strategic activity taking place in and through cyberspace. Indeed, much of what we have been observing in cyberspace concerns activity below the threshold of armed attack taking place over a long period of time.[36] Consider, for example, the Chinese cyber-enabled IP theft taking place on a large scale for over a decade.[37] Some have argued that cyber activity is part of an age-old intelligence contest.[38] Others have suggested that cyberspace enables a new dimension of power politics in which cyber campaigns could potentially become a salient means, alternative to war, for achieving strategic advantage.[39] Such an understanding of the multifaceted and simultaneous nature of cyber activity taking place below the threshold of armed attack has not emerged in the context of Locked Shields.

6. STRATEGIC SIGNALING

Third, at the strategic level, military exercises deal with the achievement of overall state objectives. Historically, military exercises have been used for geopolitical messaging. According to Clem, "*where* the exercises are conducted, *how many* personnel are involved, *what* countries they are drawn from, and the *types of weaponry* employed are all key elements in strategic positioning or, one might say, posturing."[40] Military exercises can be a means through which states can show both their willingness and capacity to operate (jointly). In this way, they can serve not only to deter potential adversaries and project power but also to reassure allies. Take the Autumn Forge

35 Ralf Bendrath, "The Cyberwar Debate: Perception and Politics in US Critical Infrastructure Protection," *Information and Security* 7 (2001): 80–103; Helen Nissenbaum, "Where Computer Security Meets National Security," *Ethics and Information Technology* 7, no. 2 (2005): 61–73; Myriam Dunn Cavelty, *Cyber-Security and Threat Politics: US Efforts to Secure the Information Age* (Abingdon: Routledge, 2008).

36 Joshua Rovner, "Cyber War as an Intelligence Contest," *War on the Rocks*, September 16, 2019, https://warontherocks.com/2019/09/cyber-war-as-an-intelligence-contest/.

37 James C. Mulvenon, "Chinese Cyber Espionage," *Testimony before the Congressional-Executive Commission on China*, June 25, 2013, https://www.cecc.gov/sites/chinacommission.house.gov/files/CECC%20Hearing%20-%20Chinese%20Hacking%20-%20James%20Mulvenon%20Written%20Statement.pdf.

38 Rovner, "Cyber War as an Intelligence Contest."

39 Michael Warner, "A Matter of Trust: Covert Action Reconsidered," *Studies in Intelligence* 63, no. 4 (2019): 33–41, https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol-63-no-4/index.html; Richard J. Harknett and Max Smeets, "Cyber Campaigns and Strategic Outcomes," *Journal of Strategic Studies* (2020). For a longer discussion on how to interpret the activity below the threshold of armed attack, see also Robert Chesney et al., "Policy Roundtable: Cyber Conflict as an Intelligence Contest," *Texas National Security Review*, September 17, 2020, https://tnsr.org/roundtable/policy-roundtable-cyber-conflict-as-an-intelligence-contest/; Robert Chesney and Max Smeets, eds., *Deter, Disrupt or Deceive? Assessing Cyber Conflict as an Intelligence Contest* (Georgetown University Press, 2022).

40 Ralph Clem, "Military Exercises as Geopolitical Messaging in the NATO-Russia Dynamic: Reassurance, Deterrence, and (In)stability," *Texas National Security Review* 2, no. 1 (November 2018), https://tnsr.org/2018/11/military-exercises-as-geopolitical-messaging-in-the-nato-russia-dynamic-reassurance-deterrence-and-instability/.

exercise series started by NATO in the mid-1970s. These exercises began with a limited goal of enhancing NATO's deterrence and defense posture but, over time, came to serve a much more subtle strategic purpose. As Palmer writes, "[The Autumn Forge exercises] became the outer face of a wider and deeper process of post-Vietnam political revitalization and military transformation of the Alliance, spear-headed by the United States, but with the growing involvement of the European Allies, which culminated in NATO's London and Washington summit meetings in 1977 and 1978, respectively."[41] In the same period (the 1970s and 1980s), the *Zapad* (West) and *Soyuz* (Alliance) exercises were introduced by the Warsaw Pact as a counterbalance.

Military exercises may not always contribute to stability through geopolitical messaging—and thus, more is not always better. Case studies have shown that they can equally invoke the classic security dilemma, in which the countries perceive these exercises as a form of aggression rather than a necessary form of defense. Bernhart and Sukin have looked at the joint military exercises held by the United States and South Korea and have found that North Korea systematically responds with aggression to these exercises.[42] This means that joint military exercises could exacerbate the danger of the security situation.

So what about the role of Locked Shields in strategic messaging? As noted, the scenario for the last Locked Shields exercise concerned an inter-state dispute between two fictional countries, Berylia and Crimsonia, including a hostile Crimsonian diaspora in Berylia and NATO involvement.[43] This scenario is not dissimilar to the current conflict situation in Ukraine. The Locked Shields organizers never stated that deterrence or other geopolitical messaging was their official aim.

It is hard to gauge the real effect Locked Shields—and CCDCOE more generally—has on Russia's cyber posture. Still, there is sufficient evidence to suggest that Russia closely monitors the exercises. In fact, one of the most comprehensive media reports on Locked Shields 2019 was written by three former Russian servicemen, explaining the setup of the scenario, how it has evolved over time, and which teams ranked highest in the contest. The influential *Sputnik News* has a page dedicated to the theme of "cyber NATO forces" with several articles about Locked Shields.[44] About the exercise in 2019, the paper has two articles entitled "Estonia Has Become a War

41 Diego A. Ruiz Palmer, "Military Exercises and Strategic Intent through the Prism of NATO's Autumn Forge Exercise Series, 1975–1989," in *Military Exercises: Political Messaging and Strategic Impact*, ed. Beatrice Heuser, Tormod Heier, and Guillaume Lasconjarias (NATO Defense College, 2018), 61-92, p. 66.

42 Jordan Bernhart and Lauren Sukin, "Joint Military Exercises and Crisis Dynamics on the Korean Peninsula," *Journal of Conflict Resolution* 65, no. 5 (2020): 855–888.

43 Michael Birnbaum, "In These Cyber War Games, The Fictional Foe Launching Attacks Sounds a Lot Like Russia," *Washington Post*, May 4, 2018, https://www.washingtonpost.com/world/europe/in-these-cyber-war-games-the-fictional-foe-launching-attacks-sounds-a-lot-like-russia/2018/05/03/06494f8c-47cb-11e8-8082-105a446d19b8_story.html.

44 *Sputnik News*, "NATO-Cyberarmy," https://ee.sputniknews.ru/trend/cyberNATO/?_ga=2.24902077.60752902.1591777423-1143849204.1591777423.

Zone—In Cyberspace" and "Locked Shields: NATO in Estonia Holds the World's Largest Cyber-Orders."[45] Both articles argue that NATO's cyber operations, and especially those taking place in the center in Tallinn, are offensive in nature and primarily directed towards Russia. Despite the nature of cyberspace, *where* military cyber exercises are conducted still seems to matter.

The latest Russian military doctrine in 2018 highlights as one of the main threats NATO member countries' "use of information and communication technologies for military and political purposes for the implementation of actions contrary to international law, directed against the sovereignty, political independence, and territorial integrity of the state."[46] In an interview for the newspaper *Red Star* (the official outlet of the Russian armed forces), Mikhail Popov, deputy secretary of the Security Council of the Russian Federation, directly linked Russia's new doctrine to the activities of the CCDCOE (as well as the NATO Strategic Communications Centre of Excellence in Riga) but not to the specific Locked Shields exercise.[47]

Finally, it is hard to promote NATO's principles of *collective* defense through the Locked Shields exercise.[48] A goal that the organizers have expressed from the very beginning of the exercises was inter-team cooperation and collaboration between the countries' Blue Teams in tackling cyber incidents. In early iterations, teams were able to score extra points for cooperation with other teams. Yet it proved difficult to measure cooperation in a meaningful way during Locked Shields and reconcile it with the competition element of the exercise. Cooperation scoring was eliminated in 2016.

45 Alexander Khrolenko, "Эстония превратилась в зону 'боевых действий.' В киберпространстве," *Sputnik News*, December 4, 2019, https://ee.sputniknews.ru/columnists/20191204/18656708/estonia-kiberuchenija-nato.html.

46 Vladimir Putin, "Военная доктрина Российской Федерации," http://static.kremlin.ru/media/events/files/41d527556bec8deb3530.pdf.

47 According to Popov, it "contributed to the emergence of a new sphere of military-political confrontation.… The NATO Strategic Communications Center in Riga, created by the alliance, and the joint center of excellence in the field of cyber defense in Tallinn [that is, the CCDCOE], are aimed at waging a large-scale information war. The main objective of these structures is to disable computer networks of critical facilities and the infrastructure of a potential adversary – as you understand, primarily Russia – by disrupting the functioning of public administration systems, financial institutions, enterprises, power plants, railway stations and airports." Anastasia Sviridova, "Совет Безопасности РФ выступает за международную стабильность на равноправных и взаимовыгодных условиях," *Red Star*, March 15, 2019, http://redstar.ru/my-znaem-i-pomnim-chto-takoe-vojna/?attempt=1.

48 A related aspect that interviewees reported was the reputational element of Locked Shields. According to Interviewee 7, while not all countries send their best teams and the organization seeks to downplay the concept of winning, Locked Shields is a way to show your allies what you are capable of in cyberspace. Winning Locked Shields gives bragging rights: "It is cool to compare yourself to others." Likewise, you do not want to end up last on the scoreboard year in, year out. Interview, Interviewee 7, 2020, Microsoft Teams.

7. FUTURE RESEARCH AND POLICY DEVELOPMENT

What purpose do the military cyber exercises serve? And what are the benefits and pitfalls of military cyber exercises? To answer these questions, this paper presents an in-depth case study of Locked Shields. The above analysis highlights several opportunities for further research and policy development.

First, future research can look more closely at the potential of different learning environments. For instance, a realistic exercise starts with a realistic "battleground." Setting up such an environment for military cyber exercises is challenging and does not come cheap. An extreme case is the US Department of Defense plans for a new global cyber training environment that is expected to cost roughly US $1 billion.[49] Other countries will not have the resources available to create a range close to this size and complexity. At the same time, there might be opportunities for international collaboration that have so far been unexplored.

Second, the collection and analysis of data also deserves more attention. As noted, unlike conventional exercises, military cyber exercises are run in a virtual environment. The benefit of this setup is the ability to collect data. Activities of different teams can be more easily recorded. While traffic logs and other data are not always easy to digest, data collection often creates the opportunity for more granular analyses of moves ex post facto.

Third, Stevens notes that "[c]yber security is peculiarly prone… to overemphasis on speed and acceleration, appropriating as it does not only the times of human others but the times of machines, the computing technologies that work at substantial fractions of the speed of light. The cyber security practices enabled by the appropriation of machine temporalities disclose the potential circumvention of customary ethics and normal political process in the names of speed and security."[50] As explained above, cyber exercises may exacerbate this bias. During an exercise lasting only a few days, such as Locked Shields, planners tend to overemphasize the importance of sudden, highly disruptive cyber attacks and underemphasize—and thus also underprepare for—the role of ongoing cyber activity below the threshold of armed attack. This suggests that future scenario development could benefit from the integration of more campaign analysis.

49 Mark Pomerleau, "Army Releases $1B Cyber Training Request," *Fifth Domain*, June 12, 2020, https://www.fifthdomain.com/dod/cybercom/2020/06/12/army-releases-1b-cyber-training-request/.

50 Tim Stevens, *Cyber Security and the Politics of Time* (Cambridge: Cambridge University Press, 2015), 16.

ACKNOWLEDGMENTS

For written comments on early drafts, the author is indebted to Henrik Beckvard, Myriam Dunn Cavelty, Andreas Hagman, Lilly Muller, Jacquelyn Schneider, Michael Widmann, and four anonymous reviewers. The author would also like to thank Brita Achberger for her excellent research assistance.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

Cyber Resilience versus Cybersecurity as Legal Aspiration

Lee A. Bygrave
Professor of Law
Norwegian Research Center for Computers and Law
Department of Private Law
University of Oslo
Oslo, Norway
lee.bygrave@jus.uio.no

Abstract: In recent years, 'cyber resilience' has sailed up as a supplement to the more traditional discourse on 'cybersecurity'. It even threatens to take over the latter as an engineering and regulatory goal. Some policy entrepreneurs believe that 'cyber resilience' rather than 'cybersecurity' ought to be a primary aim of information systems development. In their view, the quest for cybersecurity downplays or overlooks the fact that insecurity is a fundamental, inescapable element of the digital world, whereas the premise of cyber resilience is that cyber threats are the rule, not the exception; cyber resilience thereby allegedly embraces a perspective offering a more realistic approach to threat management. Proponents of cyber resilience as an overarching goal also see it as offering greater flexibility and pragmatism than the traditional concern for cybersecurity—characteristics that are especially important in a fast-changing threat environment. Yet in terms of methodology, operationalization and legal norms, to what degree does focusing on cyber resilience *actually* differ from cybersecurity-focused discourse? Are the differences more cosmetic than substantial? And to what degree is an overriding concern for resilience compatible with legal requirements, particularly those recognized in human rights jurisprudence? It is with such questions that this paper is concerned. The paper's underlying message is that cyber resilience ought not to take priority over cybersecurity as a public policy goal; rather, both goals ought to be met. This message is buttressed by four basic points. First, the interrelationship of cyber resilience and cybersecurity as conceptual constructs and public policy goals is marked by ambiguity and normative muddle, and this state of affairs has helped allow misleading characterizations of their differences to proliferate. Second, existing law significantly restricts the degree to which a quest for cyber resilience may replace the quest for cybersecurity. Third, contemporary security engineering methods together with recent legislative reforms have injected greater

flexibility and threat awareness into cybersecurity thinking. Fourth, operationalization of cyber resilience is achievable within an appropriately comprehensive 'security-by-design' framework, such as that required under EU law.

Keywords: *resilience, cyber resilience, cybersecurity, security by design, law, human rights*

1. INTRODUCTION

With roots partly in early 1970s ecology,[1] the notion of 'resilience' has since gained steady ground as a technical-systemic goal and public policy ideal across a diverse range of fields.[2] Indeed, it is described as having become a 'quasi-universal answer to problems of security and governance, from climate change to children's education, from indigenous history to disaster response, and from development to terrorism'.[3] Over the last decade, 'resilience' and 'cyber resilience' have gained increasing prominence in discourse concerned with the protection of critical infrastructure, including its cyber dimensions.[4] European lawmakers are now frequently framing legislative initiatives on the security of data, information systems, and networks in terms of enhancing the resilience of the assets concerned.[5] Cyber resilience has additionally become a central goal within military thinking.[6] This is also the case for NATO: resilience is a cornerstone of its agenda, rooted in the commitment of its members under its founding Treaty to 'maintain and develop their individual and

1 A seminal and oft-cited work being C. S. Holling, 'Resilience and stability of ecological systems', *Annual Review of Ecology and Systematics* 4, no. 1 (1973): 1–23, p. 1. Holling defines 'resilience' in terms of a capacity to persist—more specifically, the 'persistence of relationships within a system and… a measure of the ability of these systems to absorb changes of state variables, driving variables, and parameters, and still persist' (17). However, the origins of the resilience notion reach back further than the 1970s and spring from a variety of other disciplines as well: see further Peter Rogers, 'The Evolution of Resilience', *Connections: The Quarterly Journal* 19, no. 3 (2020): 13–32, https://doi.org/10.11610/connections.19.3.01.

2 See generally Jeremy Walker and Melinda Cooper, 'Genealogies of resilience: From systems ecology to the political economy of crisis adaptation', *Security Dialogue* 42, no. 2 (2011): 143–160, https://doi.org/10.1177/0967010611399616; Benoît Dupont, 'The cyber-resilience of financial institutions: significance and applicability', *Journal of Cybersecurity* 5, no. 1 (2019): 1–17, pp. 4–5, https://doi.org/10.1093/cybsec/tyz013.

3 Claudia Aradau, 'The promise of security: resilience, surprise and epistemic politics', *Resilience* 2, no. 2 (2014): 73–87, p. 73, https://doi.org/10.1080/21693293.2014.914765.

4 See e.g. U.S. Presidential Policy Directive on Critical Infrastructure Security and Resilience (February 2012); European Commission, 'Resilience, Deterrence and Defence: Building strong cybersecurity for the EU' (JOIN(2017) 450 final).

5 See e.g. European Commission, Proposal for a Directive of the European Parliament and of the Council on measures for a high common level of cybersecurity across the Union, repealing Directive (EU) 2016/1148 (COM(2020) 823 final) (NIS2 Directive; also hereinafter 'NIS2D'); European Commission, Proposal for a Regulation on digital operational resilience for the financial sector and amending Regulations (EC) No. 1060/2009, (EU) No. 648/2012, (EU) No. 600/2014 and (EU) No. 909/2014 (COM(2020)595 final).

6 See e.g. Chuck Crossett, 'Resilience in the Face of Cyberattacks: Cyber Resilience Guidance for Military Systems', *Johns Hopkins APL Technical Digest* 34, no. 4 (2019): 511–516, https://www.jhuapl.edu/Content/techdigest/pdf/V34-N04/34-04-Crossett.pdf, and references cited therein.

collective capacity to resist armed attack'.[7] In recent years, resilience has emerged as an ever more salient fixture of NATO policy documents, particularly in respect of civil-military cooperation,[8] and this extends to the cyber sphere.[9]

For many scholars, engineers, and policy entrepreneurs, resilience should have been flagged in the cyber sphere a long time ago. In their opinion, the quest for cybersecurity takes insufficient account of the fact that *in*security is a basic, inescapable condition of the digital world, whereas the premise of cyber resilience is that cyber threats are the rule rather than the exception and that information systems will inevitably be vulnerable to attack and disruption. For instance, Rothrock states:

> Resilience is… creatively pessimistic in assuming that a large number of cyberattacks will inevitably be directed against any and every organization, that security devices will inevitably fail to stop a significant fraction of those attacks, and that management's top cybersecurity priority should be reducing the volume and severity of damage and loss as well as staying in business or on mission during a breach.[10]

He continues by noting that, whereas resilience 'engages a reality' (i.e. 'that bad things are happening'), security hinges on the 'hope of evading or postponing that reality'.[11]

Similarly, Björck et al. claim:

> [T]he concept of resilience essentially treats adverse cyber events as a part of normal operations. The difference to the concept of security can therefore be crucial—it allows organizations to incorporate counter measures and contingency plans as a part of what could be considered as this new 'normal' condition.[12]

7 Article 3 of the North Atlantic Treaty ('Washington Treaty') adopted on 4 April 1949.

8 See esp. 2021 Brussels Summit Communiqué (14 June 2021) paras. 30 and 32, https://www.nato.int/cps/en/natohq/news_185000.htm, along with the Strengthened Resilience Commitment (14 June 2021), https://www.nato.int/cps/en/natohq/official_texts_185340.htm. See also Wolf-Diether Roepke and Hasit Thankey, 'Resilience: The First Line of Defence', *NATO Review* (2019), https://www.nato.int/docu/review/articles/2019/02/27/resilience-the-first-line-of-defence/index.html. All accessed 9 March 2022.

9 See esp. 2016 Warsaw Summit Cyber Defence Pledge (8 July 2016), paras. 2–5, https://www.nato.int/cps/en/natohq/official_texts_133177.htm.

10 Ray A. Rothrock, 'Digital Network Resilience: Surprising Lessons from the Maginot Line', *The Cyber Defense Review* 2, no. 3 (Fall 2017): 33–40, pp. 36–37, https://www.jstor.org/stable/26267383.

11 Ibid., 37.

12 Fredrik Björck, Martin Henkel, Janis Stirna, and Jelena Zdravkovic, 'Cyber Resilience – Fundamentals for a Definition', in *New Contributions in Information Systems and Technologies*, eds. Alvaro Rocha, Ana Maria Correia, Sandra Costanzo, and Luis Paulo Reis (Springer, 2015): 311–316, pp. 311, 315, https://doi.org/10.1007/978-3-319-16486-1_31.

This narrative has clear parallels to earlier ecology-focused elaborations of resilience.[13] It pitches cyber resilience—defined by Björck et al. as 'the ability to continuously deliver the intended outcome despite adverse cyber events'[14]—as a more realistic, flexible, and pragmatic point of departure for threat management than cybersecurity. Further, the basic goal of cyber resilience is described as quite different from that of cybersecurity: whereas the latter is allegedly concerned with 'fail-safe' measures, the former is concerned with 'safe-to-fail' measures—that is, '[r]esilient systems should be designed to be able to fail in a controlled way, rather than being designed to solely protect against failure'.[15] Rothrock complements this perspective by stating:

> Security is analogous to the 'wall' function of the Maginot Line. It is about preventing an attack.… Resilience is about standing up to do business while fighting back and recovering.[16]

Some policy entrepreneurs go so far as to argue that cyber resilience ought to receive priority over cybersecurity as a primary aim of information systems development.[17] In their view, the immense number of cybersecurity breaches and attack vectors in today's far-flung infrastructure networks necessitate a reworking of defence strategy in line with a resilience-focused mindset. Reading between the lines of this narrative, we see that cybersecurity is cast as rigid, unwieldy, outdated, and thereby doomed to being out-manoeuvred.

In this paper, I push back against the thrust of this narrative. In doing so, I revisit the relationship between security and resilience in the cyber context, showing that the nature of the relationship is muddied by conceptual ambiguity and exaggeration of the differences between the ways in which the two notions are operationalized. Moreover, I show that the above narrative overlooks the legal framework for their operationalization. Indeed, the bulk of discourse on cyber resilience fails to take adequate account of legal factors. My basic argument is that cyber resilience, while seemingly being a more nimble-footed 'creature' than cybersecurity, is not necessarily

13 Cf. Holling, 'Resilience', 21. 'A management approach based on resilience… would emphasize the need to keep options open, the need to view events in a regional rather than a local context, and the need to emphasize heterogeneity. Flowing from this would be not the presumption of sufficient knowledge, but the recognition of our ignorance; not the assumption that future events are expected, but that they will be unexpected. A resilience framework can accommodate this shift of perspective, for it does not require a precise capacity to predict the future, but only a qualitative capacity to devise systems that can absorb and accommodate future events in whatever unexpected form they may take.'

14 Björck et al., 'Cyber resilience', 311, 312.

15 Ibid., 315.

16 Rothrock, 'Digital Network Resilience', 36–37. At the risk of stating the obvious, the Maginot Line was an extensive fortification system constructed along the eastern border of France after the First World War as a bulwark against future German aggression, but which Hitler's forces deftly skirted around in the 'Blitzkrieg' of 1940. It has subsequently become a metaphor for false security.

17 See e.g. Jim Alkove, 'Cyber security is no longer enough: businesses need cyber resilience', World Economic Forum, 2021, https://www.weforum.org/agenda/2021/11/why-move-cyber-security-to-cyber-resilience/; Mike Baukes, 'Cybersecurity Is Dead', *Forbes*, 6 June 2017, https://www.forbes.com/sites/forbestechcouncil/2017/06/06/cybersecurity-is-dead/.

fundamentally different from the latter in its operationalization as a legal norm, and that this operationalization is achievable within an appropriately comprehensive legal framework for 'security by design'. Moreover, cybersecurity law and jurisprudence on fundamental human rights severely limit the degree to which a quest for supposedly 'safe-to-fail' cyber resilience may lawfully prevail over a quest for 'fail-safe' cybersecurity in the context of information systems development. This is especially so when such systems process personal data.

2. CONCEPTUAL AMBIGUITY AND VACUITY

Resilience is a notion with a slippery meaning that overlaps with a variety of other closely related notions. The latter include resistance, robustness, security, and persistence, each of which has contested meanings. Delineating the precise semantic contours of each of these notions is daunting, not least owing to their typical function as proxies for a range of 'higher order' interests or values that will tend to shape their connotations in context-dependent ways. Applying a somewhat 'tongue-in-cheek' tone, Dunn Cavelty et al. aptly describe resilience as 'the new superhero in town', albeit a superhero with an enigmatic multiplicity of forms and meaning:

> Resilience is mysterious: she can be in many places at the same time, takes on various forms, slips into different subject-bodies, and eludes clearly-defined dimensions of time—some say she is only ever emergent in essence. She is a typical postmodern heroine, existing in different universes, with various stories of origin—her multiple personalities imbue different characteristics, normative concepts, and ways of interacting with subjects.[18]

Policy documents that promote 'resilience' without properly defining the term add to the mystery. Such documents are commonplace.[19] In this regard, EU cyber policy has much to answer for, having often mentioned 'resilience' without squarely addressing its meaning. For example, the European Commission's first policy strategy on cybersecurity in 2013 made numerous references to 'resilience' and 'cyber resilience',

18 Myriam Dunn Cavelty, Mareile Kaufmann, and Kristian Søby Kristensen, 'Resilience and (in) security: Practices, subjects, temporalities', *Security Dialogue* 46, no. 1 (2015): 3–4, p. 3, https://doi.org/10.1177/0967010614559637.

19 See also George Christou, *Cybersecurity in the European Union: Resilience and Adaptability in Governance Policy* (Palgrave Macmillan, 2016), 11: 'Many cybersecurity strategies within and beyond Europe refer to developing effective cyber resilience, but without adequately defining and deconstructing what resilience is, what it looks like at different stages, and the preconditions and governance forms required to achieve it.' See also Dupont, 'The cyber-resilience of financial institutions', 8, which documents a similar lack of definition in marketing by the private sector.

but it never defined them directly.[20] The same applies to the Commission's subsequent iterations of EU cybersecurity policy in 2017[21] and 2020.[22]

The aforementioned NATO policy documents also tend to invoke 'resilience' without defining the term directly.[23] Nonetheless, they implicitly pitch the notion as a capacity to resist and as a reduction of vulnerability. Further, *NATO's Glossary of Terms and Definitions* expressly defines 'cyberspace resilience' as '[t]he overall technical and procedural ability of systems, organizations and operations to withstand cyber incidents and, where harm is caused, recover from them with no or acceptable impact on mission assurance or continuity'.[24]

By contrast, the U.S. Presidential Policy Directive on Critical Infrastructure Security and Resilience issued by the Obama administration in 2012 defined 'resilience' as 'the ability to prepare for and adapt to changing conditions and withstand and recover rapidly from disruptions'.[25] It also defined 'security' in terms of 'reducing the risk to critical infrastructure by physical means or defense cyber measures to intrusions, attacks, or the effects of natural or manmade disasters'.[26] Although rather diffuse, both definitions are better than no definitions.

Also problematic is ambiguity regarding the interrelationship of resilience and security, along with their cyber counterparts. Again, EU cyber policy has much to answer for in this regard, as its presentation of how the two notions relate is inconsistent and confusing. If we consider the EU General Data Protection Regulation (GDPR),[27] this effectively treats 'resilience' as a property of 'security',[28] whereas the Network and Information Systems Security Directive (NISD)[29] omits 'resilience' from its definition

20 European Commission, 'Cybersecurity Strategy of the European Union: An Open, Safe and Secure Cyberspace' (JOIN(2013) 1 final).

21 European Commission, 'Resilience, Deterrence and Defence: Building strong cybersecurity for the EU' (JOIN(2017) 450 final).

22 European Commission, 'The EU's Cybersecurity Strategy for the Digital Decade' (JOIN(2020) 18 final).

23 See 2021 Brussels Summit Communiqué and Strengthened Resilience Commitment and 2016 Warsaw Summit Cyber Defence Pledge.

24 *NATO Glossary of Terms and Definitions* (AAP-06 2021): 37. Cf. NATO, 'Resilience and Article 3', 11 June 2021, https://www.nato.int/cps/en/natohq/topics_132722.htm. 'Resilience is a society's ability to resist and recover from such shocks and combines both civil preparedness and military capacity.'

25 U.S. Presidential Policy Directive on Critical Infrastructure Security and Resilience. Cf. U.S. Department of State and the U.S. Department of Homeland Security's Cybersecurity and Infrastructure Security Agency, *A Guide to Critical Infrastructure Security and Resilience* (November 2019), 11.

26 Ibid.

27 Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) OJ L 119 of 4 May 2016, 1–88.

28 See the exemplifications of 'technological and organisational measures' to ensure the security of personal data in Article 32(1)(b) (referring to 'the ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services') and Article 32(1)(c) (referring to 'the ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident').

29 Directive (EU) 2016/1148 of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union, O.J.L. 194 of 9 July 2016, 1–30.

of 'security of network and information systems'[30]—the case also with the proposed NIS2D.[31] However, both the NISD and proposed NIS2D define 'security of networks and information systems' in terms of an ability 'to resist', so 'resilience' is arguably implicit in the definition. Yet, 'resistance' is not necessarily the same as 'resilience': the former may connote simply the ability to counter and defend against attack or disruption, whereas the latter connotes the ability to maintain or resume operations during or after an attack or disruption.

To complicate matters further, the Commission's proposal for a new Directive on the resilience of critical entities (CED proposal)[32] defines 'resilience' as 'the ability to prevent, resist, mitigate, absorb, accommodate to [*sic*] and recover from an incident that disrupts or has the potential to disrupt the operations of a critical entity' (Article 2(2)). This definition encompasses up-front protection—typically a hallmark of security—as well as the ability to resume operations in the event of disruption. It seemingly treats security as a property of resilience, whereas, say, the GDPR embraces the converse line. As for the Commission's NIS2D proposal, parts of it pitch resilience as a means of achieving greater cybersecurity,[33] while other parts describe the proposal as ultimately concerned with enhancing resilience.[34]

Some of these inconsistencies might be justified in light of differences in the policy objectives of the respective instruments: for instance, data protection law (of which the GDPR is a part) has traditionally been concerned more with the security of personal data than the resilience of information systems. And some of the conceptual ambiguity might be deliberate for the same sorts of reasons that 'national security' remains an underdetermined concept in EU law—that is, to preserve leeway in the delicate balancing of EU member states' sovereignty with the transnational harmonization endeavours of the EU. Yet, they come with a cost in the form of a persistent lack of coherence. All up, we face a rather tangled and confusing legal-conceptual framework for cyber resilience and cybersecurity as public policy ideals. This is nothing new. It bespeaks a continuation of the problems that have beset the evolution of EU

30 Article 4(2) NISD, which defines such security as 'the ability of network and information systems to resist, at a given level of confidence, any action that compromises the availability, authenticity, integrity or confidentiality of stored or transmitted or processed data or the related services offered by, or accessible via, those network and information systems'.

31 NIS2D proposal.

32 Proposal for a Directive of the European Parliament and of the Council on the resilience of critical entities (COM(2020) 829 final).

33 See e.g. NIS2D proposal, Article 1(1) ('This Directive lays down measures with a view to ensuring a high common level of cybersecurity within the Union') and recital 59 ('Maintaining accurate and complete databases of domain names and registration data (so called "WHOIS data") and providing lawful access to such data is essential to ensure the security, stability and resilience of the DNS, which in turn contributes to a high common level of cybersecurity within the Union').

34 See e.g. Legislative Financial Statement accompanying NIS2D proposal, 2: 'The revision's objective is to increase the level of cyber resilience of a comprehensive set of businesses operating in the European Union across all relevant sectors, to reduce inconsistencies in the resilience across the internal market in the sectors already covered by the Directive and to improve the level of joint situational awareness and the collective capability to prepare and respond.'

regulatory policy on cybersecurity over many years. As Arnbak observed of this policy prior to 2016, it has developed without a 'coherent understanding at the EU level about how to define "security", and how its underlying values operate, relate or should be interpreted',[35] a state of affairs that 'has allowed powerful actors to paint communications security any color [*sic*] they like'.[36]

This deficit in legal-conceptual stringency is not just troubling as a matter of principle or logic. It may also thwart proper understanding of cyber resilience and cybersecurity by the many stakeholders that provide vectors for attacks on information systems and connected infrastructure.[37] Taken as a whole, these stakeholders have widely varying degrees of consciousness, competence, and interest in threat management—evidenced in part by the numerous data security breaches that regularly occur.[38] Legal-conceptual muddle risks undermining clarity of the norms and methodologies for meeting the goals concerned, thereby undermining their traction. This is especially the case in respect of resilience which is 'a powerful concept but… sufficiently ambiguous that it can become counterproductive if used carelessly'.[39] Further, legal-conceptual muddle within the EU regulatory system may have ramifications for organizations outside the EU, particularly those that align their resilience and security policy efforts with EU standards. NATO is one such organization.[40]

3. SIMPLISTICITY AND EXAGGERATION

Conceptual haziness and muddle around cyber resilience, cybersecurity, and their interrelationship also help to allow misleading characterizations of their differences to proliferate. For example, in a relatively systematic attempt to classify the defining features of each notion as methodology or approach, Björck et al. claim that, in terms of architecture, resilience lays emphasis on 'multi-layered protection' with each layer

35 Axel M. Arnbak, *Securing Private Communications: Protecting Private Communications Security in EU Law – Fundamental Rights, Functional Value Chains and Market Incentives* (Wolters Kluwer, 2016), 50.

36 Ibid., 51.

37 For an overview of stakeholders, see e.g. Kjell Hausken, 'Cyber resilience in firms, organizations and societies', *Internet of Things* 11 (September 2020), section 3.2, https://doi.org/10.1016/j.iot.2020.100204.

38 See further Section 3.

39 Dupont, 'The cyber-resilience of financial institutions', 13.

40 See e.g. 2021 Brussels Summit Communiqué, esp. para. 65: 'NATO-EU cooperation has reached unprecedented levels, with tangible results in countering hybrid and cyber threats… Political dialogue between NATO and the EU remains essential to advance this cooperation. We will continue to develop and deepen our cooperation…. The current strategic environment and the COVID pandemic underscore the importance of NATO-EU cooperation in the face of current and evolving security challenges, in particular in addressing resilience issues, emerging and disruptive technologies, the security implications of climate change, disinformation, and the growing geostrategic competition.' See also e.g. 2021 Strengthened Resilience Commitment, esp. para. 10: 'As we strengthen our efforts to build resilience, we will continue to work with our partners engaged in similar efforts…. This includes the European Union, with which we will continue to build on the scope for mutually complementary and beneficial coordination in strengthening resilience, and to seek further concrete steps and effective synergies.'

offering the possibility of protection and recovery, whereas security focuses on a single layer of protection.[41] On a related note, they state:

> [A] resilience approach would have a much more profound effect on the systems being 'secured', leading to the need to let the resilience be an inner part of the IT systems and the general operation of the business. Resilience… needs to be built-in rather than an add-on.[42]

They further claim that a resilience focus is more holistic and sensitive to interconnectedness than a security focus.[43] Additionally, the former is supposedly more concerned than the latter with the environment in which a system operates, both as a source of threat and as a source of assistance in the event of recovery.[44] The implication is that a security mindset predominantly looks upon the system in isolation and sees the surrounding environment in terms of potential threat rather than potential aid.

Treated as a whole, these claims paint the possible differences between the two approaches in simplistic, exaggerated strokes. For one thing, the painting evokes each approach as relatively uniform when, in fact, they each encompass rather disparate sets of strategies and goals. To exemplify, the concern for resilience may come with differing levels of ambition. Dupont observes:

> At a fundamental level, there is some disagreement over the true meaning of resilience: for some, it entails the capacity of a system to withstand a shock and return to its original state, while for others it implies an evolutionary process leading to adaptation and a new state of equilibrium.[45]

Conceiving resilience in terms of an ability to maintain 'continuous' delivery of intended outcomes (as Björck et al. do) sets a higher level of ambition than a conception of resilience as an 'ability to prepare for and adapt to changing conditions and withstand and recover rapidly from disruptions' (to cite the aforementioned definition by the U.S. Presidential Policy Directive). The latter tolerates interruptions; the former does not. Such a difference has consequences for the ambit and stringency of the risk management involved.

More importantly, current security modelling and strategy have a greater degree of sophistication than Björck et al. indicate. Done properly, contemporary security engineering takes into account a large range of risks, vectors, and other variables, and

41 Björck et al., 'Cyber Resilience', 314–315.
42 Ibid., 314.
43 Ibid.
44 Ibid.
45 Dupont, 'The cyber-resilience of financial institutions', 2.

attempts to implement measures across a variety of operational levels.[46] The painting also omits the fact that numerous resilience measures, such as fault detection, error recovery, security renewability, backup, and fallback, are staples of modern security strategy.[47] Accordingly, security work involves resilience work as a matter of standard practice. Anderson notes in this regard:

> Providing the ability to recover from security failures, as well as from random physical and software failures, is the main purpose of the protection budget for many organisations. At a more technical level, there are significant interactions between protection and resilience mechanisms.[48]

Similarly, Bodeau et al. observe:

> As cyber resiliency techniques mature and are more widely adopted, the disciplines of cyber resiliency, cyber security, and conventional security will merge. In the meantime, there is overlap between some of the cyber resiliency techniques and some approaches used in conventional security or cyber security.[49]

Moreover, the principle of 'security by design' (SbD) increasingly informs security engineering. The SbD mantra essentially urges computer engineers and others involved in the building of information systems architecture to consider the security needs for that architecture before it is built and to integrate those needs in its subsequent design and construction.[50] Part of this objective is to ensure that information systems settings are initially configured to optimize security—an objective often referred to as 'security by default'. The thrust of SbD is to embed security within the heart of information systems architecture and, indeed, within the heart of the organization(s) deploying the systems, as opposed to a situation where security is a last-minute add-on. Thus, SbD has strong parallels to the thrust of a resilience approach as portrayed by Björck et al., along with other experts.[51] It also bears emphasizing that SbD has gone from being simply a software engineering principle to being a 'hard law' principle in both primary and secondary EU legislation—a point elaborated in Section 4.

46 See generally Ross Anderson, *Security Engineering: A Guide to Building Dependable Distributed Systems, 3rd Edition* (Wiley, 2020).

47 Ibid., section 7.3.

48 Ibid., 271–72.

49 Deborah Bodeau, Richard Graubart, William Heinbockel, and Ellen Laderman, *Cyber Resiliency Engineering Aid – The Updated Cyber Resiliency Engineering Framework and Guidance on Applying Cyber Resiliency Techniques* (MITRE Corporation, 2015), 17.

50 See generally Lee A. Bygrave, 'Security by Design: Aspirations and Regulatory Realities', *Oslo Law Review* 8, no. 3 (2021), section 2.1, forthcoming.

51 See e.g. Ron Ross, Victoria Pillitteri, Richard Graubart, Deborah Bodeau, and Rosalie McQuaid, *Developing Cyber Resilient Systems: A Systems Security Engineering Approach*, NIST Special Publication 800-160, vol. 2, rev. 1 (National Institute of Standards and Technology, 2021): 1, https://doi.org/10.6028/NIST.SP.800-160v2r1. 'Cyber-resilient systems are systems that have security measures or safeguards "built in" as a foundational part of the architecture and design and that display a high level of resiliency.'

I am not suggesting that cyber resilience and cybersecurity are fully commensurate with each other as disciplines or operational goals. Indeed, they may sometimes be in tension with each other, particularly when priority is given to ensuring the overall resilience of a communications network rather than ensuring the security of specific network nodes.[52] Nor am I suggesting that all aspects of the canvas that Björck et al. paint are incorrect. Some aspects are valid for older approaches to security. However, even when they point in a credible direction, the delineated lines need more nuance. For instance, there is little doubt that the traditional approach to information systems security has predominantly focused on safeguarding the confidentiality, integrity, and availability of the assets concerned—the classic 'CIA' triad of security properties—with relatively little concern for resilience as such (over and above safeguarding 'availability', which implies a measure of resilience).[53] We see this exemplified in the 27001 series of standards of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) on information security management.[54] To the limited degree that they tackle the issue of resilience, they are concerned more with the continuity of security management in the face of disruptions rather than the continuity of business operations more generally.[55] Nonetheless, business continuity is addressed in depth in the ISO 22301 standards,[56] which are intended to be operationalized in conjunction with the 27001 series.

As for differences in environmental and systemic perceptions, the aid potential of the surrounding environment may well be traditionally less visible through a security lens than through a resilience lens, and the latter may well be more sensitive to systemic interconnections. Nonetheless, security as an exercise in state-of-the-art risk management involves carefully distinguishing between various aspects of the environment depending on their threat levels and, in the process of doing so, recognizing those aspects that are relatively benign if not helpful for the assets concerned.[57] Further, a security lens is not necessarily significantly restricted in its ability to take account of systemic interconnections. Admittedly, certain legal cybersecurity regimes, such as that of the EU, have developed in an incremental, atomistic way, focusing on particular sectors, value chains, or threats while sidelining others.[58] This has resulted

52 As exemplified by the core architecture for the early Internet which was constructed with simplicity, flexibility, openness, and resilience predominantly in mind, not with security. See e.g. Bruce Schneier, *Click Here to Kill Everybody* (W.W. Norton and Company, 2018), 22–24 and references cited therein.

53 See also Ross et al., *Developing Cyber Resilient Systems*, 80.

54 ISO/IEC 27001:2013 (Information technology – Security techniques – Information security management systems – Requirements), https://www.iso.org/standard/54534.html.

55 Ibid., Appendix A.17.

56 ISO/IEC 22301:2019 (Security and resilience – Business continuity management systems – Requirements), https://www.iso.or/standard/75106.html.

57 See e.g. IT Security Association Germany (TeleTrusT) and ENISA, 'IT Security Act (Germany) and EU General Data Protection Regulation: Guideline "State of the Art" – Technical and Organisational Measures' (version 1.9_2021-09 EN): 74. 'Risk management consists of systematic risk assessment and identification, monitoring and handling of risk areas. The goal is to systematically identify *opportunities* and risks to a company and to assess these risks with reference to the likelihood they will occur and to their quantitative effects on company values.' Emphasis added.

58 See generally Arnbak, *Securing Private Communications*; Christou, *Cybersecurity in the European Union*.

in a number of regulatory 'silos'. Yet, this development has not been primarily a function of myopic traits intrinsic to a security focus. Rather, it has been a function of extraneous political and logistical factors, such as the aforementioned challenges involved in managing the tension between nation states wishing to determine security norms for themselves and the needs of international bodies to develop harmonized transnational norms that may undercut national sovereignty interests.[59]

Finally, it bears emphasizing that many of the cybersecurity breach scandals of recent years have not been due to a fault in security methodology as such; their immediate cause has been a failure to implement proper security measures in the first place.[60] Various economic factors disincentivizing the implementation of such measures have exacerbated these failures.[61] So, too, have weaknesses in the application of legal sanctions.[62]

4. EUROPEAN LAW AND FUNDAMENTAL RIGHTS

The European legal framework for cybersecurity is on the way to becoming considerably more comprehensive in scope and holistic in approach. It is thereby attempting to bridge many of the regulatory 'silos' that have hitherto existed, such as those of, respectively, product safety and critical infrastructure. A prominent example in point is the NIS2D proposal that dispenses with the current Directive's focus on, and distinction between, operators of essential services and digital service providers.[63] Other examples include the European Commission's flagging of the interdependence between 'cyber' and 'non-cyber' threat dimensions, particularly in respect of critical infrastructure,[64] and the Commission's proposed extension of European Health and Safety Requirements for machinery products to embrace cybersecurity risks stemming from malicious third party actions.[65] This bridging process is also melding security

59 Ibid.

60 Consider e.g. the 2017 'Equifax hack', which led to the unauthorized disclosure of sensitive information on over 148 million persons, mostly in North America. The poor security measures Equifax allegedly had in place at the time of the attack (as documented in the class action initially brought before the U.S. District Court for the Northern District of Georgia Atlanta Division (Civil Action File No. 17-CV-3463-TWT)) are almost beyond belief. See further, Stanford Law School, http://securities.stanford.edu/filings-documents/1063/EI00_15/2019128_r01x_17CV03463.pdf.

61 See e.g. Ross Anderson and Tyler Moore, 'The Economics of Information Security', Science 314 (2016): 610–613; Hadi Asghari, Michel van Eeten, and Johannes M. Bauer, 'Economics of Cybersecurity', in *Handbook on the Economics of the Internet*, eds. Johannes M. Bauer and Michael (Edward Elgar, 2016), 262–287.

62 See e.g. NIS2D proposal, 5 ('Member States have been very reluctant to apply penalties to entities failing to put in place security requirements or report incidents').

63 See Article 2 and Annexes 1 and 2 NIS2D proposal.

64 See CED proposal.

65 European Commission, Proposal for a Regulation of the European Parliament and of the Council on machinery products (COM(2021) 202), Annex III, with amended 'European Health and Safety Requirement' (EHSR) 1.2.1 on safety and reliability of control systems: 'Control systems shall be designed and constructed in such a way that: (a) they can withstand, where appropriate to the circumstances and the risks, the intended operating stresses and intended and unintended external influences, including malicious attempts from third parties to create a hazardous situation.'

and resilience measures. Fresh examples in point are the Commission's legislative proposals concerning, respectively, an Artificial Intelligence Act (AIA)[66] and a Regulation on resilience for the financial sector.[67]

Further, recent legislation and legislative initiatives promote more explicitly an approach to security that is adaptable, iterative, and ongoing. This agenda runs counter to the perception of cybersecurity as locking in a rigid, one-off set of measures amounting, in effect, to a digital equivalent of the Maginot Line. For instance, the EU Cybersecurity Act (CA)[68] states: 'Security should be ensured throughout the lifetime of the ICT product, ICT service or ICT process by design and development processes that *constantly evolve* to reduce the risk of harm from malicious exploitation' (recital 12; emphasis added). It also recognizes the threat-infused nature of information systems' environs, stating that protection of the security of ICT products, ICT services, or ICT processes should be 'to the highest possible degree, in such a way that the occurrence of cyberattacks *is presumed* and their impact is anticipated and minimised' (recital 12: emphasis added). To take another example, the GDPR requires—as one of the 'technical and organisational measures' for security under Article 32—a 'process for regularly testing, assessing and evaluating the effectiveness' of these measures (Article 32(1)(d)). This is reinforced by the references to 'state of the art' in the incipits of Articles 32(1) and 25(1) GDPR, which necessitate periodic revisiting of security practices in light of changing perceptions as to which technical and organizational standards provide optimal security.[69] The same can be said of Articles 14(1) and 16(1) NISD, which also use 'state of the art' as a touchstone.

Another important ingredient in the reform of European cybersecurity law and policy is the notion of 'security by design'. In respect of EU law, SbD has gone from being merely a technical engineering standard to becoming entrenched as a hard law norm and, in relation to the processing of personal data, a full-fledged regulatory principle inhering not just in secondary legislation but also in the EU's constitutional fabric.[70] It has done so hand-in-hand with the related mantras 'data protection by design and

66 Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021) 206 final). See Article 15(1): 'High-risk AI systems shall be designed and developed in such a way that they achieve, in the light of their intended purpose, an appropriate level of accuracy, robustness and cybersecurity, and perform consistently in those respects throughout their lifecycle.'

67 Proposal for a Regulation on digital operational resilience for the financial sector and amending Regulations (EC) No. 1060/2009, (EU) No. 648/2012, (EU) No. 600/2014 and (EU) No. 909/2014 (COM(2020)595 final). See esp. Article 8(2): 'Financial entities shall design, procure and implement ICT security strategies, policies, procedures, protocols and tools that aim at, in particular, ensuring the resilience, continuity and availability of ICT systems, and maintaining high standards of security, confidentiality and integrity of data, whether at rest, in use or in transit.'

68 Regulation (EU) 2019/881 of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act) O.J.L. 151 of 7 June 2019, 15–69.

69 See generally IT Security Association Germany and ENISA, 'IT Security Act'.

70 Bygrave, 'Security by Design', section 3.

by default' (DPbDD) and, to a more limited degree, 'privacy by design' (PbD).[71] Indeed, SbD-related norms initially emerged in EU secondary legislation on data protection,[72] and then spread to other EU secondary legislation, such as the CA,[73] NISD,[74] Electronic Communications Code,[75] and Medical Devices Regulation,[76] as well as current legislative proposals, such as the AIA. As elaborated further below, these norms have also found anchorage in the European Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) and Charter of Fundamental Rights of the European Union (CFREU).

This trend is part of a burgeoning 'by design' discourse aimed at integrating various values into technology production processes. Its core rationale is a conviction that embedding or 'hardwiring' the values into such processes will substantially enhance their traction.[77] Admittedly, some of the legislative manifestations of this trend are criticized for resulting in fluffy, vacuous, and enigmatic provisions—the case in particular with Article 25 GDPR, which concerns, on its face, DPbDD but also implicitly embraces SbD.[78] Nonetheless, their 'by design' dimension manifests an intention of lawmakers that the norms be more than mere window dressing. The European Data Protection Board (EDPB) rightly observes that '[e]ffectiveness is at

71 See further Lee A. Bygrave, 'Data Protection by Design and by Default: Deciphering the EU's Legislative Requirements', *Oslo Law Review* 4, no. 2 (2017): 105–120, p. 105, https://doi.org/10.18261/issn.2387-3299-2017-02-03; Line Jasmontaite, Irene Kamara, Gabriela Zanfir-Fortuna, and Stefano Leucci, 'Data protection by design and by default: framing guiding principles into legal obligations in the GDPR', *European Data Protection Law Review* 4, no. 2 (2018): 168–189, p. 168, https://doi.org/10.21552/edpl/2018/2/7.

72 Initially in Article 17(1) and recital 46 of Directive 95/46/EC of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, O.J.L. 281 of 23 November 1995, 31–50 (repealed); thereafter in Article 4(1) of Directive 2002/58/EC of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector, O.J.L. 201 of 31 July 2002, 37–47. See now Articles 32(1), 25, and 5(1)(f) GDPR, together with Articles 29 and 20 of Directive (EU) 2016/680 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, O.J.L. 119 of 4 May 2016, 89–131, together with Articles 33 and 27 of Regulation (EU) 2018/1725 of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No. 45/2001 and Decision No. 1247/2002/EC, O.J.L. 295 of 21 November 2018, 39–98.

73 See e.g. Article 51(i).

74 See esp. Articles 14(1) and 16(1).

75 Directive (EU) 2018/1972 of 11 December 2018 establishing the European Electronic Communications Code, O.J.L. 321 of 17 December 2018, 36–214. See particularly Article 40(1) and recitals 94, 95, and 97.

76 Regulation (EU) 2017/745 of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No. 178/2002 and Regulation (EC) No. 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC, O.J.L. 117 of 5 May 2017, 1–175. See esp. Article 10(1) in conjunction with paragraph 17(2) of Annex I to the Regulation.

77 See further e.g. Lee A. Bygrave, 'Hardwiring Privacy', in *The Oxford Handbook of Law, Regulation, and Technology*, eds. Roger Brownsword, Eloise Scotford, and Karen Yeung (Oxford University Press, 2017), 754, 755.

78 See e.g. Ari Waldman, 'Data Protection by Design? A Critique of Article 25 of the GDPR', *Cornell International Law Journal* 53 (2020): 147–167, p. 148; Aurelia Tamò-Larrieux, *Designing for Privacy and its Legal Framework: Data Protection by Design and Default for the Internet of Things* (Springer, 2018), 209.

the heart of the concept of data protection by design'.[79] Thus, adds the Board, the paramount objective of DPbDD is 'the *effective implementation* of the principles and *protection* of the rights of data subjects into the appropriate measures of the processing'.[80] With contextual modifications, this observation rings true also for other provisions in which SbD-related ideals are manifest. In other words, the fundamental legislative agenda here concerns yet another instance of attempted bridging: this time it is the attempted bridging of law in books with law in practice.[81]

The latter point becomes more obvious when we also consider the way in which SbD (and DPbDD) ideals have become rooted in fundamental human rights. The main catalyst for this development is the jurisprudence of the Court of Justice of the EU (CJEU) and the European Court of Human Rights (ECtHR), both of which have injected SbD-related ideals into Europe's constitutional framework for fundamental rights protection. The ECtHR judgment in *I v. Finland* (2008)[82] is seminal in this respect. In that case, the Court held that Finland violated its positive obligation to ensure respect for private life under Article 8 ECHR because of a failure to provide 'practical and effective protection to exclude *any possibility* of unauthorized access' to patient data at a public hospital.[83] Although omitting explicit reference to SbD or DPbDD, the thrust of the judgment implies an approach in line with their ideals. In effect, the judgment renders SbD (and, concomitantly, DPbDD) an essential requirement of a state's positive obligations to secure respect for the right laid out in Article 8 ECHR. It also imposes a high threshold for meeting that requirement, at least in relation to ensuring the confidentiality of data concerning a person's health. There are solid grounds for viewing these obligations as extending to other types of personal data and to other functionalities than just maintaining data confidentiality.[84]

The CJEU is likely to take a similar approach regarding the obligations flowing from Articles 7 and 8 CFREU,[85] along with Article 16 of the Treaty on the Functioning of the EU (TFEU).[86] This is partly because of the so-called homogeneity clause in Article 52(3) CFREU,[87] together with the Court's confirmation that Article 7 CFREU

79 EDPB, 'Guidelines 4/2019 on Article 25: Data Protection by Design and by Default' (version 2.0; 20 October 2020), para. 13.

80 Ibid., para. 96. See also Jasmontaite and others, 'Data protection', 176–177 (noting that Article 25 GDPR introduces an 'obligation by result' and not of 'process').

81 See also Bygrave, 'Security by Design', section 5.2.

82 Appl. No. 20511/03.

83 Ibid., para. 47. Emphasis added. In a judgment handed down shortly afterwards, the Court again emphasized the need to secure 'practical and effective protection' of a person's right under Article 8(1) in respect of internet-related conduct, although the case did not deal directly with cybersecurity issues. See *KU v. Finland*, Appl. No. 2872/02 (2008), esp. para. 49.

84 See further Arnbak, *Securing Private Communications*, 81; Bygrave, 'Data Protection by Design', 111.

85 Article 7 CFREU lays down a right to respect for private life similar to Article 8 ECHR, whereas Article 8 CFREU stipulates a right to the protection of personal data.

86 Article 16 TFEU also lays down a right to data protection, albeit one to be primarily respected by EU institutions.

87 Article 52(3) basically states that rights in the Charter shall have the same meaning and ambit as corresponding rights in the ECHR, though without preventing European Union law from providing more extensive protection than under the ECHR.

'must therefore be given the same meaning and the same scope as Article 8(1) of the ECHR, as interpreted by the case-law of the European Court of Human Rights'.[88] Moreover, the CJEU has strongly implied that the 'essence' of the right to data protection laid down in Article 8 CFREU requires respect for 'certain principles of data protection and data security', meaning that 'Member States are to ensure that appropriate technical and organisational measures are adopted against accidental or unlawful destruction, accidental loss or alteration of the data'.[89]

Obviously, this jurisprudence—along with the aforementioned cybersecurity norms laid out in secondary legislation—has consequences for any attempt to ditch or significantly scale back cybersecurity efforts in the name of cyber resilience. In particular, it prevents easy replacement of attempts to institute reasonable 'fail-safe' security measures by putatively 'safe-to-fail' resilience measures, especially in respect of personal data.

One must also bear in mind that, in the longer term, cybersecurity as such might become explicitly recognized as a fundamental right in itself and in a way that extends beyond data protection.[90] This could happen, for example, as an extension of the 'right to liberty and security of person' under Article 6 CFREU.[91] Such a development is still some way off, but it would cement cybersecurity even stronger in the European legal order were it to eventuate.

5. CONCLUSION

In this paper, I have explored the interrelationship of 'cyber resilience' and 'cybersecurity' with a view to questioning claims that the former is, relative to the latter, a fundamentally different and superior goal for information systems development. I have done so taking account of the European legal landscape in which these goals are to be operationalized. The paper shows that the differences are not as fundamental as some policy entrepreneurs seem to assume, especially in light of recent reforms of cybersecurity law which seek to introduce greater flexibility and threat awareness into security thinking.

When recalibrating defence strategies in order to enhance their ability to cater for 'fast moving reality'—to use the phrasing for the thematic thrust of CyCon 2022—one

88 Case C-400/10 PPU, *McB v. LE*, judgment of 5 October 2010 (ECLI:EU:C:2010:582), para. 53.

89 Joined Cases C-293/12 and C-594/12, *Digital Rights Ireland Ltd v. Minister for Communications, Marine and Natural Resources and Others and Kärntner Landesregierung and Others*, judgment of 8 April 2014 (Grand Chamber) (ECLI:EU:C:2014:238), para. 40.

90 Vagelis Papakonstantinou, 'Cybersecurity as *praxis* and as a *state*: The EU law path towards acknowledgement of a new right to cybersecurity?', *Computer Law & Security Review* 44, no. 1 (2022), https://doi.org/10.1016/j.clsr.2022.105653.

91 For discussion, see Luca Tosoni, 'The Fundamental Right to (Cyber) Security: A Critical Appraisal of Article 6 CFREU' (Ph.D. diss., University of Oslo, 2022), forthcoming.

must be careful not to throw, as it were, the baby out with the bath water or, at the risk of mixing metaphors, to throw the tortoise out with the pond water. One must also remember that appearances can deceive. The outwardly slow-moving but inwardly industrious tortoise in Aesop's well-known fable ended up being able to reach the finish line before the hare. Similarly, cybersecurity-focused strategies are reasonably capable of reaching their intended 'finishing line' *if* they are suitably conceived and implemented, and they may do so as equally well if not better than a supposedly more nimble-footed resilience-focused strategy. I emphasize 'if' in the previous sentence because cybersecurity strategies come in many shapes and sizes, as do resilience strategies. The important point is that a cybersecurity strategy designed and deployed in accordance with state-of-the-art risk management is not the equivalent of a Maginot Line.

By encouraging an approach to cybersecurity that is holistic, flexible, adaptable, iterative, and effective, recent legislative reforms in Europe are attempting to improve the practical bite of legal security rules. These reforms are undoubtedly inspired by resilience-focused ideals from which springs, to use the words of Kaufmann, a 'self-made, emergent, and strictly temporary notion of security'.[92] In a sense, the reforms are trying to provide the security tortoise with greater power and agility. An alternative and more preferable view is that they are attempting to fuse the advantages of the tortoise with the advantages of the hare. For security and resilience must ultimately go hand in hand.

ACKNOWLEDGEMENTS

Work on this paper was conducted under the aegis of the research project 'Security in Internet Governance and Networks: Analysing the Law' (SIGNAL), funded by the Research Council of Norway and UNINETT Norid AS (grant number 247947), the research project 'Vulnerability in the Robot Society' (VIROS), funded by the Research Council of Norway (grant number 288285), and the research project 'Governance of Health Data in Cyberspace' (CyberHealth), funded by Nordforsk (grant number 81105). I am particularly grateful to Arild Jansen, Tobias Mahler, and Ingrid Winther for helpful commentary on a previous draft of the paper.

92 Mareile Kaufmann, 'Resilience governance and ecosystemic space: a critical perspective on the EU approach to Internet security', *Environment and Planning D: Society and Space* 33 (2015): 512, 524, https://doi.org/10.1177/0263775815594309.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

Public-Private Partnerships and Collective Cyber Defence

John Morgan Salomon
Regional Director, Continental Europe, Middle East, and Africa
FS-ISAC (Financial Services Information Sharing and Analysis Center)
US/UK/SG/International
jsalomon@fsisac.com

Abstract: The line between traditional military conflict and cyberattacks and related electronic crimes is increasingly blurred. Attacks by a wide range of malicious actors, ranging from state-sponsored or -sanctioned groups to independent criminal gangs of all sizes and levels of maturity, can degrade the confidentiality, integrity, and availability of entities critical to the continued orderly functioning of democratic free-market societies – and thus the continued orderly functioning of the global economy. To defend ourselves, we must prepare for, detect, defend against, and learn from such threats. How can the public and private sectors cooperate to this end?

This paper explores how private sector firms can work with regulators, law enforcement, national cybersecurity, and defence agencies to help protect civil society. It presents the lessons learned by the FS-ISAC over more than 20 years of global activity, building collective defence capability in the financial sector and developing structures and capabilities to mitigate risks quickly and effectively through cooperation with government partners.

The paper also lists recommendations for government stakeholders, from policymakers to defence and police agencies, on how to effectively support, and work with, private industry across national borders. Activities such as regulation, exercises, crisis response frameworks, and trust networks are all part of the repertoire that government and industry are jointly developing in the 21st century to safeguard against an ever-evolving and increasingly broad cyber threat landscape.

Keywords: *public-private partnership, resilience, information sharing, collective defence, cyber threat intelligence*

1. INTRODUCTION AND DISCLAIMER

Private industry plays a significant part in every country's prosperity and stability; some sectors are more vital than others to continued peaceful order. When individual companies or whole sectors are threatened or attacked by cyber actors, it can compromise the functioning of an entire society.

This paper addresses four major areas of concern:

1. From the perspective of the private sector and those it employs to defend it against cyber threats, it rarely matters what motivates or who is responsible for a cyberattack.
2. Public-private cooperation is vital for preventing, identifying, defending against, and following up on cyber-attacks.
3. National defence agencies must play a significant role in protecting against and responding to cyberattacks.
4. Compared with other types of public-private partnerships, cooperation between private-sector and defence entities is weak to non-existent in many countries and should be dramatically improved.

It is written from the perspective of private-sector cybersecurity rather than that of formal military warfighting experience, and it aims to reconcile and combine the knowledge and resources of the private sector with the capabilities and powers of government agencies as part of holistic cyber defence. Preparing for, and defending against, cyber threats and attacks are topics affecting entire industries and societies. It also requires ever-closer coordination between all stakeholders – potential victims, subject matter experts, government agencies, and entities that can coordinate information flows and actions between all of these.

Cooperation involves pragmatic rules and structures of collective preparation and defence. It relies on the creation and adoption of norms, relationships, and the fundamental mentality of working together across both private industry and the public sector. We cannot completely stop cyberattacks that threaten the stability and continuity of society, but we can dramatically reduce their effectiveness and attractiveness as a way for any malicious actors, regardless of nature or objectives, to degrade the robustness and continuity of economies and the societies they serve.

Many examples cited in this paper concern the global financial sector. Because of a mix of visibility, regulatory burden, resources, and other factors, the financial industry is among the most developed in terms of information security maturity and cooperation.

Thus, it is a valuable source of evidence and experience on how to strengthen our collective cyber defence abilities.

A note about terminology: in the private sector information-security field, the term 'attack' tends to be used much more loosely than in defence circles.[1] It is any threat action that can compromise one or more elements of the 'CIA triad' of confidentiality, integrity, and availability. It does not require that there be a state actor or any specific type of objective or target. Furthermore, 'private sector' is used very broadly, to apply to all entities not dedicated to public service (i.e., it may include publicly-owned commercial firms, as well as non-governmental not-for-profit entities).

2. CURRENT SITUATION – 30,000 YEARS OF WARFARE ON HALF A PAGE

Armed confrontation has, unfortunately, been part of human civilization throughout history. War has affected entire nations to varying degrees over the centuries. At one extreme is highly ritualized combat between designated warriors on a field of battle, with limited collateral damage. At the other lies total war, with entire populations massacred by vast armies or enslaved by conquering legionaries, and entire cities and national infrastructure laid waste by modern weapons of incredible power.

Likewise, asymmetric and proxy warfare have existed in various forms for millennia. The Second World War saw conventional warfare between militarized nations rise to a peak of intensity and sophistication. Since then, asymmetric warfare has increasingly become the norm for how wars are fought. Guerrilla warfare and terrorism directly target civilian populations and economic infrastructure and seek to destabilize political processes.

In developed, wealthy countries, societal and international norms surrounding conflict have fundamentally shifted. As civilian societies and economic activity have become more interdependent and connected, the willingness to mobilize a nation's entire economic and social resources for war, or to completely smash those of an opponent, has decreased.

Rival countries have always jockeyed for competitive advantage outside of military combat. In recent decades, this has involved increasingly sophisticated techniques in all areas of digital information technology. Communication tools and techniques, such as propaganda, disinformation, and other means of destabilising opposing societies,

1 The term 'attack' is defined in 'Cyber Attack', *Glossary*, National Institute of Standards and Technology, U.S. Department of Commerce, Computer Security Resource Center, https://csrc.nist.gov/glossary/term/cyber_attack.

have thus become much more relevant to how nations compete off the battlefield during nominal peacetime.

Fake news on social media that easily reaches millions of consumers, espionage and data leaks, information theft from individuals and companies, destructive malware capable of knocking out multinational firms,[2] and digital sabotage campaigns able to cause major outages in vital infrastructure such as power grids and water supply are all various aspects of this.

So what are 'war' and 'conflict' today? This paper argues that we must expand our understanding of the terms. Over the past two or three decades, a new dimension has been added: ongoing background threats and attacks targeting information resources and related infrastructure have become 'business as usual'. This is not a repeat of the Cold War in anything beyond a weak metaphorical sense, since it does not involve purely state actors trying to gain an edge while attempting to avoid a direct shooting war. It is a major evolution of hybrid, low-intensity hostilities, which includes constant cyberattacks and, as such, is a new phenomenon requiring new paradigms of cooperation between all stakeholders in affected societies.

3. ACTORS, TYPES OF ATTACKS, AND MOTIVATION

Many simplistic models list types of actors posing a significant risk for entire economic sectors – mainly state actors, organized cybercrime gangs, nihilistic saboteurs, and activists.[3] Among other criteria, actors can be differentiated by their level of knowledge, resources, and motivation.

2 One example is the 2017 NotPetya/WannaCry impact on Maersk, Merck, and Mondelez.

3 We will leave out informal/amateur attackers (e.g., 'script kiddies'), insiders, or inadvertent actors, as these do not tend to pose a systemic threat.

Similarly, cyber threats and attacks come in many forms. Table I provides some examples.[4]

TABLE I: A PARTIAL LIST OF TYPES OF CYBERATTACKS AND SOME OF THEIR ATTRIBUTES

Type	Techniques	Immediate goal(s)	Example(s)
Active offensive, destructive operations	Sabotage through destructive malware Denial of Service	Economic/business disruption	Stuxnet 2007 distributed denial of service (DDoS) attack against Estonia Fancy Bear / APT28
Information theft	Espionage Exfiltration through malware/spyware	Competitive advantage Network/systems reconnaissance Reputational damage	LightBasin attacks on telcos OilRig/APT34
Fraud, monetary theft	Business Email Compromise (BEC) Phishing / spear phishing Transaction fraud	Operational financing Destruction of confidence/credibility	2016 SWIFT/Bangladesh incident
Disinformation and economic propaganda	Organized troll farms[5] Government media attacks Astroturfing	Societal destabilization Political pressure from lobbying by affected organizations Reputational damage to key economic actors	Foreign-affiliated interference in US elections, Ukrainian politics, Chinese pressure on Western firms re Taiwan, Uighurs

All these types of cyberattacks have become a regular occurrence. Almost all types of attacks have increased over the past five years.[6] The European Union Agency for Cybersecurity (ENISA), in its 2021 Threat Landscape report,[7] identifies not only an increase in the sophistication and variety of tactics, techniques, and procedures (TTPs) but also a growth in malicious state actors relying on offensive security tools and false-flag attacks. The report also notes a significant rise and evolution in state actors using 'information operations as a tool to pursue states' strategic goals'.

4. ARE WE AT WAR?

This question is flawed and incomplete, and even irrelevant, from the perspective of private-sector information security staff and victims.

To those affected by malicious activity and incidents, the motivations and techniques behind cyberattacks matter little. In terms of legality and impact on victims, cybercrime

4 Obviously, this list is far from complete. Techniques are those that can pose a system-wide threat.
5 One example is the Internet Research Agency: https://www.defenseone.com/threats/2018/02/what-internet-research-agency/146085/.
6 *FBI Internet Crime Report 2020*, https://www.ic3.gov/Media/PDF/AnnualReport/2020_IC3Report.pdf.
7 *ENISA Threat Landscape 2021*, https://www.enisa.europa.eu/publications/enisa-threat-landscape-2021.

is cybercrime. At an operational level, an attack is an attack, whoever is behind it. Preparation, defence, and recovery will generally be the same.

Lately, state actors have increasingly been involved in, or responsible for, major cyberattacks, fraud/theft, and espionage against major commercial organizations. Prominent examples of this are alleged Russian campaigns, beginning with the 2007 nationwide DDoS against Estonia, espionage and sabotage operations by reported Chinese state actors from APT1 onwards, the OilRig group widely associated with Iran, and suspected North Korean involvement in the Lazarus Group. Nor are such tactics limited to rogue states,[8] whatever their goals may be.

State actors are far from the only sources of cyberattacks. The line separating purely financially or ideologically motivated groups from government-run initiatives has blurred. In some cases, such as the DarkSide and REvil groups,[9] it is likely that non-government actors may be tolerated by host governments, so long as they respect certain rules of engagement – for example, not attacking 'friendly' targets. Other actors may be politically agnostic in nature or focused on certain classes of targets, regardless of national or political affiliation – such as LulzSec,[10] Lizard Squad,[11] and Armada Collective.[12]

The nature of attackers and their motivations matters to some degree. For example, attacks by terrorist or state-affiliated cybercriminals may more readily trigger a supportive response by friendly government entities, such as national cybersecurity centres or law enforcement. The TTPs used in cyberattacks may also depend on which group is involved and can thus influence defensive measures and countermeasures.

Ultimately, the damage and cost to victims are the same, regardless of the attacker's identity. Loss of CIA of data and systems degrades normal operations. Causes include financial loss from business downtime or outright monetary theft, as well as reduced confidence in institutions and in the trustworthiness of technological mechanisms from data breaches. Victims, at least in the short term, generally do not care why or by whom they were attacked; they care about the immediate consequences to themselves or to their organization.

These examples point to a growing understanding that, whatever the actors and motivations behind cyberattacks, threats and attacks to the continued orderly

8 E.g., Operation 'Olympic Games' and the Stuxnet worm that targeted Iranian nuclear facilities. See David E. Sanger, 'Obama Order Sped up Wave of Cyberattacks against Iran', *New York Times*, 1 June 2012.

9 Robyn Dixon and Ellen Nakashima, 'Russia Arrests 14 Alleged Members of REvil Ransomware Gang, Including Hacker U.S. Says Conducted Colonial Pipeline Attack', *Washington Post*, 14 January 2022.

10 Charles Arthur, 'LulzSec: What They Did, Who They Were and How They Were Caught', *Guardian*, 16 May 2013.

11 Brian Krebs has numerous resources about Lizard Squad on his blog 'Krebs on Security', https://krebsonsecurity.com/tag/lizard-squad/.

12 Matthew Prince, 'Empty DDoS Threats: Meet the Armada Collective', Cloudflare blog, 25 April 2016.

functioning of society, government, and the economy cannot be clearly defined as either 'war' or 'not war'.

5. MANAGING RISK THROUGH PREPAREDNESS

Incident and crisis response are established concepts. Many organizations have mature business continuity, incident response, and contingency planning organizations and processes. These are increasingly supported by the realization of business leaders that understanding and reducing cybersecurity risks are not discrete goals. They are vital to the continued operation and survival of the business itself.

Although the idea of integrating cyber risk and business risk management structures only started gaining widespread traction in the past decade,[13] this trend has accelerated as boards and management better accept the need to quantitatively understand cyber risk exposure. Tools, processes, and standards allow a number to be assigned to risk levels by measuring security control implementation effectiveness, thus greatly simplifying this process.

As key businesses become more digitized and thus more susceptible to cyber threats, regulators have begun to pay closer attention to corporate cyber risk practices, as evidenced by proposed regulations such as the EU Digital Operational Resilience Act (DORA).[14] This has resulted in more measurable, evidence-driven investment in cybersecurity controls and thus better preparedness.

Well-designed and regularly tested business continuity management (BCM) and incident response (IR) are part of any mature organizational cybersecurity risk management structure. But on its own, even a highly resilient company with well-resourced, proven, strong information security defences and recovery proficiency is unlikely to invest in what could be construed as protecting commercial competitors.

As a result, there is little inherent incentive for profit-oriented firms, or any other entities with limited resources, to join forces. Company policies may even explicitly prohibit cybersecurity-related information sharing with external organizations. Those organizations that are the most mature, active, and cooperative are those that have succeeded in measuring cybersecurity risk[15] and connecting control effectiveness to other measures of organizational and strategic risk.[16] They can show clear value from

13 An example of an exposition of this problem is provided by Brian Contos, 'Close the Gap Between Cyber-Risk and Business Risk', Dark Reading, 2019, https://www.darkreading.com/risk/close-the-gap-between-cyber-risk-and-business-risk.

14 European Commission, Proposal for a Regulation of the European Parliament and of the Council on Digital Operational Resilience for the Financial Sector and Amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014 and (EU) No 909/2014, 24 September 2020, COM/2020/595 final.

15 This can happen through resources like those offered by FAIR. See https://www.fairinstitute.org/.

16 The Operational Risk eXchange's cyber risk taxonomy is an interesting approach to this. See https://managingrisktogether.orx.org/operational-risk-taxonomy.

strong information security and resilience maturity. Unfortunately, many firms still struggle with this.

6. ENTER CNI

Critical national infrastructure (CNI) is a comparatively new concept, first gaining significant traction in its current form in the late 1990s. The idea of CNI extends the concepts of managing risk to critical components at an organizational level, to entire sectors and societies.

Government organizations and rules, such as the 1998 United States Presidential Decision Directive/NSC-63[17] defining the idea of a sector-specific ISAC (information sharing and analysis centre), and the establishment of the UK National Infrastructure Security Co-ordination Centre in 1999 (now the Centre for the Protection of National Infrastructure, or CPNI), established the importance of certain economic sectors as part of national security and stability and created frameworks and resources to help protect these from cyberattacks.

Agencies that are best placed to retaliate against threat actors increasingly cooperate with each other to protect their country. The placement of the Dutch National Cyber Security Centre (NCSC)[18] within the Ministry of Justice and Security, and the 2008 establishment of the US National Cyber Investigative Joint Task Force (NCIJTF)[19] across the CIA, Departments of Defense and Homeland Security, and National Security Agency are examples of growing inter-agency cooperation and communication.

Such agencies are also increasingly willing to speak out about threats. According to the ENISA 2021 Threat Landscape Report, government agencies such as the US Department of Justice, NCSC-UK, and the European Council have increasingly issued public statements and/or sanctions related to state-affiliated or -sponsored threat actors. This points to a growing willingness to inform and include citizens and private sector audiences as a part of disruption activities.

17 'Presidential Decision Directive/NSC-63', May 22, 1998, https://irp.fas.org/offdocs/pdd/pdd-63.htm. An ISAC combines a central entity that coordinates industry risk reduction and resilience building efforts, as well as collecting, analysing, and disseminating information and intelligence, with a community of participating member organizations from the same economic sector, committed to sharing information about observed threats and other relevant topics that serve to augment the sector's overall defensive posture. ISAC is a rough concept rather than a strict set of criteria; numerous ISACs exist around the world for diverse CNI sectors, ranging from very informal, local groups with occasional meetings to highly mature and well-resourced global organizations with a wide range of operational services, platforms, and products.

18 National Cyber Security Centre, https://english.ncsc.nl/.

19 FBI National Cyber Investigative Joint Task Force, https://www.fbi.gov/investigate/cyber/national-cyber-investigative-joint-task-force.

Building on such developments, a growing collection of national cybersecurity agencies and affiliated public-private initiatives, such as the German UP-KRITIS cooperation group established via the 2005 national plan for the protection of information infrastructures,[20] have matured and formalized the interaction between government agencies and major private sector entities.

7. THE NEW 'CITIZEN MILITIA'

Since the Second World War, many Western democracies have abandoned conscription-based militaries. Modern military forces tend to be more professional, smaller, and more technically expert than their mid- or early 20th century counterparts. This may be because of the decline in conventional warfare between 'rich' states, even though, according to the United Nations, 'Globally, the absolute number of war deaths has been declining since 1946. And yet, conflict and violence are currently on the rise, with many conflicts today waged between non-state actors such as political militias, criminal, and international terrorist groups.'[21] Whether this is due to a historical cycle, as argued by Nassim Taleb and Pasquale Cirillo,[22] or the result of a trend towards more peaceful civilization overall[23] is beyond the scope of this paper – time will tell.

In my subjective opinion, the decline in the threat of violent conflict in most developed, wealthy countries has resulted in the average citizen becoming disassociated from the idea of war. Even when a country's armed forces are involved in a physical conflict somewhere in the world, the actions are mostly of limited scope, and most militaries in wealthy nations are relatively small, focused, and professional. Thus, with a few major exceptions, such as the US-led coalitions that invaded Iraq in 1991 and Afghanistan in 2001, our awareness of conflict is generally far removed from the widespread popular displays of patriotism and even militarism that many countries witnessed in the first half of the twentieth century.

Meanwhile, I assert that the increase in cyberattacks from various sources, in all fields of society, has resulted in a dramatic but rarely explicitly stated evolution of the idea of the 'citizen soldier'. Because of the universality of cyberattacks and the aforementioned blurring of clear distinctions between targets, both individual citizens and organizations are increasingly involved in defending their societies against such attacks. Both are victimized by a wide range of attackers; companies in particular

20 'Umsetzungsplan KRITIS des Nationalen Plans zum Schutz der Informationsinfrastrukturen', https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/umsetzungsplan-kritis.html.

21 United Nations, 'A New Era of Conflict and Violence', https://www.un.org/en/un75/new-era-conflict-and-violence.

22 Nassim Nicholas Taleb and Pasquale Cirillo, 'On the Statistical Properties and Tail Risk of Violent Conflicts' (2015), https://doi.org/10.48550/arXiv.1505.04722.

23 E.g., Stephen Pinker, *The Better Angels of Our Nature* (Penguin Press, 2010).

are increasingly engaged in leveraging in-house cybersecurity resources as part of collective defence.

Despite the aforementioned inherent lack of incentive for individual firms to cooperate, companies do react to a) regulatory pressure and b) industry good practice – it becomes much easier for cybersecurity professionals to justify investing time in community activities if either a regulator explicitly requires information sharing or a group of peer organizations is already doing so successfully and visibly. The Brazilian private security superintendent's Circular 638 of July 2021[24] is a prime example of a very aggressive regulatory mandate for information sharing.

This is even more the case when there is some sort of central entity that builds and coordinates collaboration forums and resources. The Financial Services ISAC (FS-ISAC),[25] founded in the US in 1998; the FI-ISAC,[26] founded in the Netherlands in 2003; and the Nordic Financial CERT,[27] founded in Norway in 2012, are strong examples of formal, structured coordination bodies, in this case specific to the financial sector. At a cross-sector level, the United States National Council of ISACs,[28] formed in 2003, provides coordination across multiple sector-specific entities.

Even when regulatory sharing and cooperation rules are not strong, guidance from public sector entities can help less-mature and -resourced industry sectors work together effectively. The European Network and Infrastructure Security Agency's ISAC in a Box toolkit[29] builds on shared experiences across multiple sectors to guide the establishment of formal arrangements.

Such guidance has led to an adoption of the principles of collective defence among sectors that may not be immediately obvious targets for cyberattackers but whose members nonetheless suffer greatly from various threats and are stronger together. The ENISA-led Empowering EU-ISACs[30] initiative, which provides support and coordination for various European sector-specific ISACs, demonstrates how even 'small' stakeholder groups can band together and benefit from such guidance and play an active part in collective cyber defence.

Even when not formally part of national/sector cyber defence and resilience structures, participants in collective cyber defence comprise a new paradigm of a 'citizen militia',

24 *Circular SUSEP No 638, de 27 de Julho de 2021*, Ministério da Economia/Superintendência de Seguros Privados, 3 August 2021, https://www.in.gov.br/en/web/dou/-/circular-susep-n-638-de-27-de-julho-de-2021-335760591.

25 Financial Sector Information Sharing and Analysis Center, https://www.fsisac.com/.

26 *Publiek-Private Samenwerking*, Betaalvereniging Nederland, https://www.betaalvereniging.nl/veiligheid/publiek-private-samenwerking/.

27 Nordic Financial CERT, https://www.nfcert.org/.

28 US National Council of ISACs, https://www.nationalisacs.org/.

29 'EU Agency for Cybersecurity Launches ISAC in a Box Toolkit', ENISA, 26 October 2020, https://www.enisa.europa.eu/news/enisa-news/isac-in-a-box.

30 'Empowering EU ISACs', https://www.isacs.eu/.

helping to defend their societies against potentially systemically destructive attacks of all types.

8. PUBLIC-PRIVATE PARTNERSHIPS

Public-private partnerships (PPPs) are nothing new. Numerous PPPs already exist and are constantly evolving; information security sharing and cooperation communities naturally gravitate towards PPPs to leverage the strengths of the public sector. Their value is increasingly recognized by entities such as ENISA and its counterparts elsewhere through good practices guides.[31] However, the existing global PPP ecosystem has a massive gap: systematic cooperation between private industry and entities tasked with national defence against cyberattacks is weak and must be improved.

An important distinction between government and private sector entities is that they often operate under different legal constraints and rules of engagement and have different stakeholders. Public sector entities are legally entitled to pursue criminal investigations and prosecute threat actors. Frequently, they see a 'bigger picture' and have access to knowledge and resources that are not available to groups of cooperating firms – but they also have public service obligations and more severe restrictions concerning confidential information.

Meanwhile, most private sector groups also have significant rules and norms concerning confidentiality and trust. For example, many groups prohibit the sharing of indicators of compromise or breach-related data with law enforcement or regulatory agencies, as fear of being fined or otherwise penalized can impact willingness to share with peers at all. And in a system under the rule of law, unauthorized non-governmental actors must not engage in vigilante retribution against attackers.

Thus, while the private sector cannot contribute significantly to, say, active disruption of cyber-threat actors and their resources, arrests, or state-level active countermeasures, it is a major resource for detection, analysis, and investigation and for leveraging the authority and knowledge of public sector counterparts to help better secure itself and the society it operates in. PPPs are a vehicle for this sharing of information and capabilities.

Cooperation between government entities and private sector organizations can be roughly categorized as either 'informational' or 'interactive'. Informational PPPs allow

31 *Good Practice Guide on Cooperative Models for Effective PPPs*, ENISA, 1 October 2011, https://www.enisa.europa.eu/publications/good-practice-guide-on-cooperative-models-for-effective-ppps; *Public Private Partnerships (PPPs)*, Cooperative Models, ENISA, 14 February 2018, https://www.enisa.europa.eu/publications/public-private-partnerships-ppp-cooperative-models.

the sharing of incidents and information with government entities – e.g., websites for reporting fraud and cyberattacks to law enforcement or government CSIRTs (computer security incident response teams). Conversely, government agencies maintain channels for communicating alerts and best practices with civil society. A US Cybersecurity and Infrastructure Security Agency alert titled 'Understanding and Mitigating Russian State-Sponsored Cyber Threats to U.S. Critical Infrastructure'[32] is a good example of such an alert, while the US Federal Financial Institutions Examination Council (FFIEC) Cybersecurity Assessment Tool[33] is a publicly available resource for any interested party to evaluate its own cybersecurity posture.

Interactive PPPs are best described as forums and groups where public- and private sector entities actively and regularly exchange information or otherwise jointly enhance overall defensive capability and maturity. This takes place at all levels – from strategic coordination between senior stakeholders to tactical, technical cooperation by operations staff. Good examples of these are the Europol European Cybercrime Centre's various sector advisory groups[34] and the US Financial Services Sector Coordinating Council (FSSCC).[35] Joint operations centres, such as the UK National Cybersecurity Centre's Industry 100 Programme and the US National Cybersecurity and Communications Integration Center run by the Cyber and Infrastructure Security Agency (CISA), as well as the standing representation of the Dutch Payments Agency at the Netherlands NCSC, physically co-locate operational security staff from intelligence services, cybersecurity bodies, law enforcement, and private sector firms in a 'constant forum' where they can work together rapidly and efficiently.

Such partnerships and information flows ensure that all parties benefit from each other's strengths and capabilities. During systemic cyberattacks such as 2012–2013's Operation Ababil,[36] 2020's Trickbot botnet disruption,[37] and the 2019–2022 series of critical vulnerabilities (including SolarWinds, Accellion, Microsoft Exchange, and Log4J), private sector analysis, information collection, and rapid, flexible response, combined with government agency coordination and information freely provided to industry actors, and in some cases, the takedown and prosecution of bad actors, allowed for the significant reduction of potential damage to many sectors. The standing communication and interaction facilitated by joint security centres like those listed above, and regular government agency representation at private sector events and working groups, has also repeatedly accelerated and enhanced preparedness for, and the response to, critical incidents and threats.

32 *Understanding and Mitigating Russian State-Sponsored Cyber Threats to U.S. Critical Infrastructure*, Cybersecurity and Infrastructure Security Agency, 11 January 2022.
33 *FFIEC Cybersecurity Assessment Tool*, FFIEC, May 2017.
34 Europol EC3 Partners list, https://www.europol.europa.eu/about-europol/european-cybercrime-centre-ec3/ec3-partners.
35 Financial Services Sector Coordinating Council, https://fsscc.org/.
36 Nicole Perlroth and Quentin Hardy, 'Bank Hacking Was the Work of Iranians, Officials Say', *New York Times*, 8 January 2013.
37 'Trickbot Disrupted', Microsoft, 12 October 2020.

The distinction between the two types of PPPs is not always clear-cut; sector-focused government entities such as the Israeli Finance Cyber and Continuity Centre (FC3)[38] within the Israeli national CERT (CERT-IL) are primarily clearinghouses for cyber threat information and alerts. They disseminate these to, and collect them from, the domestic sector. FC3 also organizes events and other collective action. Elsewhere, industry-specific ISACs similarly arrange information flows or just bring stakeholder entities to the same table.

It is as difficult to empirically prove the positive impact of these initiatives as it is to prove the effectiveness of any cybersecurity control at preventing cyberattacks. However, using the example of the financial sector's response to the aforementioned critical vulnerabilities, I maintain that the speed and efficiency of vulnerability and threat identification, the communication and implementation of collective protective measures, and the rapid follow-up and lessons learned contributed significantly to the relatively small amount of damage suffered by the industry compared to the technological severity of recent supply chain cyber threats and -vulnerabilities. That is sufficient evidence of their value.

9. THE ROLE OF REGULATORS AND CENTRAL BANKS – RULES, GOOD PRACTICE, AND SECTOR RISK MANAGEMENT

The 'public' side of PPPs is not limited to law enforcement or cybersecurity agencies. Regulatory[39] agencies care about ensuring stability and predictability in their national industry and are vital in forcing (or encouraging) private sector entities to work together, or to work with public sector counterparts.

The aforementioned Brazilian insurance sector Circular 638 regulation obliging information sharing, and others like it, demand both incident reporting and information exchange. The proposed European Network and Information Security 2 (NIS2) directive[40] includes provisions for risk management practices, increased incident reporting requirements, and information sharing. Even when a communication from a government agency with regulatory responsibilities does not include an unambiguous requirement to participate in a PPP or other information-sharing arrangement, it can provide significant impetus for private sector organizations to do so. The 2014 US

38 Israel Ministry of Finance Cyber and Finance Continuity Center, https://www.gov.il/en/Departments/General/cyber_center_and_financial_continuity.

39 A distinction is frequently made between 'regulatory' agencies (those who make rules) and 'supervisory' agencies (those who ensure compliance with rules). Frequently, the two functions are merged; for simplicity's sake, I use 'regulators' to refer to both.

40 'The NIS2 Directive: A High Common Level of Cybersecurity in the EU', European Parliament Think Tank, 1 December 2021.

FFIEC Cybersecurity Assessment Observations[41] and its recommendations led to a dramatic rise, particularly among small US financial services organizations, in sector information-sharing arrangements.

Regulators are also very well placed to provide guidance and good practice information. The Central Banks of Kuwait and Jordan published, respectively, the comprehensive *Cybersecurity Framework for Kuwaiti Banking Sector*[42] in 2020 and *Cybersecurity Framework for Jordan Financial Sector*[43] in 2021, which are typical of such resources. Similarly, the Monetary Authority of Singapore's *Guidelines on Risk Management Practices: Technology Risk*,[44] updated in 2021, is a useful tool and reference guide for information security organizations.

Lastly, regulators and other government agencies focused on specific industry verticals can be excellent drivers for the establishment of national public-private cybersecurity partnerships. Even when national CSIRTs or cybersecurity entities already exist, a regulator can ensure that the cybersecurity needs of 'its' sector are met. In the case of the Israeli FC3, although it is administratively part of the Israeli national CERT (CERT-IL),[45] it falls under the governance of the Israel Ministry of Finance as part of that ministry's support for sector resilience. This structural detail is important; it defines clear responsibility for cyber intelligence support for the financial sector and sends a clear signal that the cybersecurity needs of the sector are being addressed by a dedicated operational entity and are taken seriously by a responsible government agency.

10. PREPARATION, TESTING, AND EXERCISES

As Prussian Field Marshal Helmuth von Moltke (the Elder) put it, 'No plan of operations reaches with any certainty beyond the first encounter with the enemy's main force'[46] – or, more commonly, 'no plan survives contact with the enemy'.[47] Nonetheless, readiness planning is a vital part of preparing for cyberattacks.

Even though it is usually impossible to accurately predict the future, just defining responsibilities, processes, resources, expectations, limitations, competencies, and

41 'FFIEC Releases Cybersecurity Assessment Observations, Recommends Participation in Financial Services Information Sharing and Analysis Center', FFIEC, 3 November 2014.

42 *Cybersecurity Framework for Kuwaiti Banking Sector*, Central Bank of Kuwait, 20 April 2020, http://www.iefpedia.com/english/wp-content/uploads/2020/02/CSF-Feb-2020.pdf.

43 *Cybersecurity Framework for Jordan Financial Sector*, Central Bank of Jordan Financial Cyber Emergency Response Team, July 2021.

44 *Guidelines on Risk Management Practices: Technology Risk*, Monetary Authority of Singapore, 18 January 2021.

45 Although FC3 falls administratively under CERT-IL and shares overall premises, it has separate facilities, staff, and tools; data flows are segregated from the rest of the CERT.

46 Helmuth von Moltke, *Kriegsgeschichtliche Einzelschriften* (1880s).

47 Also, as Murphy's Second Law states, 'Murphy was an optimist.'

contact points is an incredibly valuable tool for all actors involved in cyber crisis response. The FS-ISAC All Hazards Framework[48] is such a resource; it is effectively a 'cookbook' for creating national, regional, or sector response plans and playbooks. Likewise, the UK Cross Market Operational Resilience Group (CMORG)[49] has as its objectives to:

- identify risks to the resilience of the financial sector;
- develop solutions to improve the operational resilience of the sector;
- share knowledge.

To ensure that contingency planning works, and to identify and remediate gaps, crisis response must be tested and exercised. The Bank of England Prudential Regulation Authority's CBEST[50] framework and its European Union counterpart framework for Threat Intelligence-based Ethical Red Teaming (TIBER)[51] both provide clear rules, processes, and structures for assessing the cyber resilience of firms based on observed real-world threats.

Exercises help ensure that crisis response resources at all levels – from tactical, technical incident response to strategic executive-level decision-making – are ready in case of a serious incident. There is no easy way to achieve total incident response preparedness, whether at a company, sector, or national/regional level. It is important that such capability is planned based on real-world requirements and realistic expectations and that participants and components are subjected to regular exercises to ensure reasonably smooth functioning rather than panic in case of a serious incident.

It follows that PPPs and CNI coordination resources benefit significantly from regular joint readiness testing exercises. Singapore's Raffles[52] exercise, the Quantum Dawn[53] series of exercises in the United States, ENISA's Cyber Europe,[54] and the Bank of England's simulation exercise (SIMEX)[55] all convene key stakeholders to identify areas for improvement in how participants prepare for, identify, respond to, and recover from cyberattacks.

The natural next step is to incorporate private sector representatives' expertise, requirements, and attack surface into actual defence preparedness activities. The NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE)[56] in Estonia

48 All Hazards Framework, https://www.fsisac.com/resources/allhazardsframework.
49 *Operational Resilience of the Financial Sector*, Bank of England, 20 October 2021.
50 *CBEST Threat Intelligence-Led Assessment*, Bank of England Prudential Regulatory Authority, January 2021.
51 *What is TIBER-EU?* European Central Bank.
52 'Singapore's Financial Sector Wraps Up Two-Day Exercise to Strengthen Business and Operational Resilience against Cyber Threats', Monetary Authority of Singapore, 22 November 2019.
53 'Fact Sheet: Quantum Dawn VI', SIFMA.
54 ENISA Cyber Europe, https://www.enisa.europa.eu/topics/cyber-exercises/cyber-europe-programme.
55 *Sector Simulation Exercise: SIMEX 2018 Report*, Bank of England, 27 September 2019.
56 Collective Cyber Defence Centre of Excellence, https://ccdcoe.org/.

has organized the annual Locked Shields exercise[57] since 2010. The exercise, while remaining primarily a military activity, has increasingly involved elements of CNI, such as industrial control systems (ICS), water infrastructure, and the power grid. In 2021, numerous financial firms participated on behalf of the financial industry as an entire sector, with FS-ISAC providing coordination[58] and support.

Increasing awareness of the importance of joint exercises between defence, other government entities, and CNI means that such public-private cyber defence crisis response, communication, and coordination exercises will continue to grow in popularity over the coming years.

11. PITFALLS, DEAD ENDS, AND CAVEATS

There are some considerations when developing sector-wide, regional, or public-private cyber defence, which various types of organizations should pay attention to.

In particular, regulators must be cautious about being too prescriptive. Rules can create unnecessary overhead, especially for smaller firms. Requirements should be actionable and based on realistic (rather than purely hypothetical) risks. Their development should ideally involve consultation with industry federations and other representatives, as well as subject matter experts.

Development and maintenance of trust are also vital. Public- and private-sector professionals can have widely differing attitudes about dealing with attacks and threats, and sometimes incompatible views of how to work with their counterparts on the other side of the public-private divide. This can take numerous forms, such as the following:

- *Defence and national security agencies* want to avoid anything, including secret information disclosure or involvement of unqualified and authorized personnel in national defence activities, which might harm national security interests. This may include cooperation with foreign counterparts or foreign-headquartered companies.
- *Law enforcement* must maintain the integrity of investigations, and thus prosecutions, by avoiding the corruption of evidence, alerting of suspects, and infringement on the right of presumption of innocence.
- *Private sector firms* worry that incidents and breaches they report may be held against them as evidence of not following recommended or mandated risk management practices, resulting in fines or reputational loss.

57 CCDCOE Locked Shields, https://ccdcoe.org/exercises/locked-shields/.
58 'FS-ISAC Leads Financial Sector in World's Largest International Live-Fire Cyber Exercise', FS-ISAC, 15 April 2021.

- *Regulators* care about sector stability and predictability, while at the same time wanting to ensure that their constituents follow applicable rules for information security practices.

It would be a counterproductive breach of trust to name specific examples where these issues have impeded joint action. Suffice it to say that they are frequent obstacles that many industry sharing and collective resilience initiatives are familiar with.

There is no substitute for a) clear and sensible rules and norms and b) trust. Both require time and resource investment all around. Trust and information sharing in particular cannot be forced, for example, by implementing a portal or platform; events, joint exercises, and statements of willingness by corporate leaders to share information are important tools in establishing cooperation.

'Regulatory harmonization' is a frequent topic of interest among private sector entities, particularly those with multinational/multijurisdictional presence. These are obliged to conform to a wide variety of requirements for cybersecurity, risk management, and information sharing. They must also do so while respecting restrictions that are sometimes at odds with other countries' rules.

Conflict or contradiction with other regulations is also a consideration – guidance on how to interpret these should be clear and unambiguous. The European Union's 2016 General Data Protection Regulation (GDPR)[59] is generally seen as a positive clarification and synchronization of data privacy rules and citizens' rights and contains clear 'carve-outs' for cybersecurity- and fraud-relevant information sharing in the interest of organizations defending themselves, their customers, and their societies.

GDPR nonetheless caused confusion and uncertainty. Many firms were unclear on what types of cyber threat information were permissible to share, and with whom. Since corporate legal and compliance departments tend to err on the side of caution, many firms issued blanket restrictions on information sharing. Without clear judicial precedent or regulatory guidance, even though numerous papers have expressed the view that such cyber threat intelligence (CTI) information sharing is a 'legitimate interest',[60] reluctance by compliance teams to countenance information sharing due to regulatory concerns will continue to be a challenge.

59 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), *Official Journal* L 119, 4 May 2016, 1–88.

60 Two excellent examples of this are Livinius Obiora Nweke and Stephen Wolthusen, *Legal Issues Related to Cyber Threat Information Sharing among Private Entities for Critical Infrastructure Protection* (2020); and Richard M. Borden, Joshua A. Mooney, Mark Taylor, and Matthew Sharkey, *Threat Information Sharing and GDPR: A Lawful Activity That Protects Personal Data*, FS-ISAC (2018).

Even within jurisdictions, there is sometimes little clarity about the interplay of related regulations. A good example of this is concern over potential conflicts[61] with the EU's GDPR and Payment Services Directive (PSD2).[62] Regulators must thus ensure that their rules and guidance remain as practical and actionable as possible to ensure they help constituent firms genuinely increase their and their sectors' security, without causing a disproportionate amount of bureaucracy or confusion.

Like everything else, cyber defence costs money. Few leaders, whether in commercial organizations or public service, will invest time and funds in collective defence without a resulting clear, quantifiable reduction of financial loss. Fewer companies yet know how to demonstrate any increased revenue or other concrete business utility from such activity. The value of contributing to collective defence networks, and in turn benefiting from shared resources and competencies, which is self-evident to most experienced information security and risk management professionals, is very difficult to quantify.

The good news is that none of these challenges are new; many have already been solved in one way or another. This paper lists numerous examples of effective private sector cooperation with law enforcement agencies; of good support from national cybersecurity organizations for their private industry counterparts; of trust structures, good practices, and joint resilience frameworks and activities that have demonstrated value in action. We can only hope that we, the private sector, and our stakeholders can continue to expand this kind of teamwork to our colleagues in the military.

12. CONCLUSION – A CALL TO ACTION

The specific arguments and recommendations of this paper are as follows:

1. Cyberattacks and related threats increasingly affect all areas of society.
2. Attacks come from a wide range of actors and groups – state-affiliated and otherwise.
3. The traditional concept of 'war' no longer applies to the constant cyberattacks that are a part of daily life for individuals and both government- and private sector organizations.
4. Critical national infrastructure is an integral part of any society's cyberattack surface and an important resource to help that society defend itself.
5. Military and other government agencies benefit significantly from cooperation with private sector entities in preparing for, and defending against, cyberattacks – and vice versa.

61 E.g., 'European Union: How Does PSD2 Interplay with GDPR', *Global Compliance News*, 23 October 2020. Although this is not specifically about cyber threat-related information, it does illustrate issues in reconciling different, highly complex regulations dealing with sensitive data processing.

62 *Payment Services (PSD 2) – Directive (EU) 2015/2366*, European Commission.

6. The collective capabilities of the private sector in facing information threats constitute the modern-day equivalent of a militia.
7. The government provides valuable leadership and encouragement for collective cyber defence activities when the private sector does not create these on its own.
8. Trust and norms are vital to effective collective defence and take time and dedication to develop and maintain.
9. Rules and regulations should make sense and avoid being excessively complicated and onerous.
10. All such capabilities require clear rules and norms and constant testing and exercising to remain effective.

There is no absolute security; we will probably never see the end of major data breaches, cyber fraud and theft, or other abuses of information assets. Threat actors and the means at their disposal will continue to evolve, as will the technologies and resources that they can attack, destroy, steal, sabotage, and otherwise disrupt.

At the same time, our defences, and our ability to strike back at perpetrators, are constantly improving. While the density, complexity, and viciousness of cyberattacks and their perpetrators is growing each year, the ability of both government and industry to not only identify and respond to these but to do so more effectively has kept track.

Vigilance, cooperation, and adaptability are the most important qualities in ensuring that our societies, citizens, and the firms and governments that serve them remain stable and safe from cyber threats.

And finally, of course, be nice to your cybersecurity professionals everywhere. They deserve it.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

Obnoxious Deterrence

Martin C. Libicki
Mary Ellen and Richard Keyser Distinguished
Visiting Professor of Cybersecurity Studies
United States Naval Academy
Annapolis, MD, United States
libicki@usna.edu

Abstract: The reigning U.S. paradigm for deterring malicious cyberspace activity carried out by or condoned by other countries is to levy penalties on them. The results have been disappointing. There is little evidence of the permanent reduction of such activity, and the narrative behind the paradigm presupposes a U.S./allied posture that assumes the morally superior role of judge upon whom also falls the burden of proof—a posture not accepted but nevertheless exploited by other countries. In this paper, we explore an alternative paradigm, obnoxious deterrence, in which the United States itself carries out malicious cyberspace activity that is used as a bargaining chip to persuade others to abandon objectionable cyberspace activity. We then analyze the necessary characteristics of this malicious cyberspace activity, which is generated only to be traded off. It turns out that two fundamental criteria—that the activity be sufficiently obnoxious to induce bargaining but be insufficiently valuable to allow it to be traded away—may greatly reduce the feasibility of such a ploy. Even if symmetric agreements are easier to enforce than pseudo-symmetric agreements (e.g., the Xi-Obama agreement of 2015) or asymmetric red lines (e.g., the Biden demand that Russia not condone its citizens hacking U.S. critical infrastructure), when violations occur, many of today's problems recur. We then evaluate the practical consequences of this approach, one that is superficially attractive.

Keywords: *cyberattack, cyberespionage, deterrence, compellence*

1. INTRODUCTION

Should the United States[1] try to modulate the cyberspace behavior of other states through a standard deterrence policy—"if you do this, I will do that"? Or, instead, should it try obnoxious deterrence? In other words, should it engage in behavior that other countries would dislike in the hope that they would agree to behave better in order to get the United States to back off?[2]

The motivation for considering obnoxious deterrence is that standard deterrence does not seem to be working very well. There is little evidence that the increasing wave of sanctions and indictments is effectively altering the cyberspace behavior of, say, Russian, Iranian, and North Korean threat actors. Granted, China, under U.S. pressure, did sharply cut back some of its cyberspace behavior (notably, economically motivated cyberespionage) in September 2015. But it partially reverted to form in early 2017 (albeit more discretely and discreetly). Discussions with them on proper behavior in cyberspace are not taking place.[3]

Advocates of standard deterrence could counter that it has not been tried in any serious manner. Indictments do little against hackers who are safe in their own country, and sanctioned entities often lack sufficient overseas assets to suffer real pain. Thus, anything other than meeting cyber with cyber (especially if it involves military force) is no true test. Raising the punishment, to be sure, ought to increase a posture's deterrent effect. Yet it may also strike others as disproportionate and could lead to counter-retaliation or escalation. Problems with proving attribution to a world audience, as well as clearly distinguishing acceptable from unacceptable behavior, would remain. The constant novelty of malicious cyberspace operations presents a further problem: the 2016 hack of the Democratic National Committee (DNC) was likened to an act of war,[4] but in a meeting of very knowledgeable individuals who tried to draw up a list of cyberspace operations that would cross a red line, nothing like the DNC hack was mentioned.

1 Although this paper has been written from a U.S. perspective, it does not argue in favor of the United States alone carrying out such cyberspace operations. Indeed, given the importance of allies to U.S. national security strategy, it is quite possible that the United States will bargain to reduce unwanted cyberspace operations against its allies as well. But offensive cyberspace operations, in large part because of their great secrecy, tend to be national, not alliance, efforts. And the focus of this paper is on operational challenges that look similar whether the United States or NATO takes the lead.

2 Unwanted cyberspace behavior comes in several forms, such as cyberattacks (cyberspace operations that compromise integrity and availability), cyberespionage (cyberspace operations that compromise only confidentiality), and cybercrime (criminal activities enabled by cyberspace operations). For our purposes, the exact nature of the unwanted cyberspace behavior is of secondary importance; it suffices that the target is willing to bear costs and take risks to end it. Mixed strategies, such as using standard deterrence to persuade other countries to discourage cybercrime and obnoxious deterrence to persuade other countries to leave infrastructures alone, are not excluded. We prefer to focus on a straight comparison.

3 Josh Rogin, "A Shadow War in Space Is Heating up Fast," *Washington Post*, November 30, 2021, https://www.washingtonpost.com/opinions/2021/11/30/space-race-china-david-thompson/.

4 Morgan Chalfant, "Former DNC Chair: Russian Election Hacking an 'Act of War,'" *The Hill*, March 29, 2017, https://thehill.com/policy/cybersecurity/326350-former-dnc-chair-calls-russian-election-hacking-an-act-of-war.

Standard cyber deterrence is also hobbled by asymmetry. The United States and the United Kingdom (whose efforts are often supported by allies) are nearly alone in having, or even trying to have, a cyber deterrence policy (Israel is a special case[5]). Russia[6] and China have not expressed an intention to create one. The U.S. narrative speaks in terms of "responsible" state actors—which we presumably are and they presumably are not. But while it has called out some actions as clearly irresponsible, the distinction between responsible and irresponsible can be unclear and, in some cases, lies less in what was done and more in how, notably when it comes to cyberespionage.[7] Recent examples include the SolarWinds operation by Russia, which resulted in sanctions, or the Hafnium/Microsoft Exchange Server operation by China, which resulted in its being called out by the United States and its allies. The United States thus looks as if it is assuming a moral superiority of the sort that a state legal system assumes over an accused criminal. The other asymmetry, one particularly relevant to cyberspace operations, is deniability. The fact of a cyberspace operation can sometimes be ambiguous, and authorship is typically denied. By contrast, punishment is usually visible and its authorship is often stated outright. The ambiguities of cyberspace favor attackers. Successful deterrence, once cyberspace operations recur, requires that one party alter its behavior; its refusal to do so or the lack of evidence that it did so is evidence that deterrence has failed. Ironically, this puts the deterrer in the position of supplicant, while the other side can simply do nothing different and clothe itself in self-righteousness for having stood up to bullying.

2. OBNOXIOUS DETERRENCE AS AN ALTERNATIVE POLICY

Obnoxious deterrence works differently, in large part, because it is more symmetric. One side, motivated by the desire to modulate another side's activity, starts doing something that the latter objects to, shifting the onus to the other side to initiate or at least participate in negotiations that establish rules of the road. Such rules bind both sides and would have each give up a practice that it formerly engaged in. If cheating occurs, the cheater can dispute the characterization (that it happened and that it violated the agreement) and/or deny the attribution—but the deterrer can respond

5 The best example is Israel's responding to a (failed) cyberattack on its water works with a (successful) cyberattack on an Iranian port. Joby Warrick and Ellen Nakashima, "Officials: Israel Linked to a Disruptive Cyberattack on Iranian Port Facility," *Washington Post*, May 18, 2020, https://www.washingtonpost.com/national-security/officials-israel-linked-to-a-disruptive-cyberattack-on-iranian-port-facility/2020/05/18/9d1da866-9942-11ea-89fd-28fb313d1886_story.html.

6 Russia's notion of "strategic deterrence" is included in its information security strategic documents, http://www.scrf.gov.ru/security/information/document5/; see also Janne Hakala and Jazlyn Melnychuk, *Russia's Strategy in Cyberspace* (Riga: NATO StratCom COE, June 2021), https://stratcomcoe.org/cuploads/pfiles/Nato-Cyber-Report_15-06-2021.pdf.

7 For an example of such a distinction, see Perri Adams, Dave Aitel, George Perkovich, and J. D. Work, "Responsible Cyber Offense," *Lawfare* [blog], August 2, 2021, https://www.lawfareblog.com/responsible-cyber-offense.

by reverting to obnoxious behavior without necessarily having to prove anything. The impact of ambiguity is more symmetric.

Obnoxious deterrence may look like compellence,[8] but that is because the line between deterrence (making something not happen) and compellence (making something happen) can be difficult to establish when trying to stop recurrent behavior. Is the idea to stop the next forbidden act, even if previous attempts to stop it failed? If so, then it looks like deterrence. Or is the idea to alter a behavior that results in repeated misdeeds? If so, it looks like compellence. The issue arises from trying to define a baseline: deterrence applies if the next misdeed would bring punishment, and compellence applies if the next misdeed is expected and thus stopping it would be a change. With nuclear deterrence, of course, deterrence was always deterrence, because there was no pattern of recurrent nuclear use.[9] But in cyberspace, recurrence is expected—and the lower the threshold by which cyberspace operations are judged to be over the line, the more frequent recurrence is. A deterrence policy founded in a tightening of thresholds may well look like compellence to the other side, because it asks for the cessation of what may have been de facto normalized behavior.

Conversely, obnoxious deterrence is not what Lucas Kello[10] has called "punctuated deterrence," which would have the defender wait until the combined cost of several successive cyberspace operations crosses a threshold before levying punishment. This version of deterrence is designed to deal with the problem that no retaliation is small enough to be proportionate when individual incidents are small—so the target waits until the cumulative costs of incidents cross some threshold. NATO, at its 2021 Brussels Summit, held that "the impact of significant malicious cumulative cyber activities might, in certain circumstances, be considered as amounting to an armed attack."[11]

Obnoxious deterrence is also not a supercharged version of Persistent Engagement,[12] which reportedly was used to stymie Russia's Internet Research Agency's attempts to interfere with the 2018 congressional elections,[13] put Russia's electric grid at risk,[14]

8 For the original definition, see Thomas Schelling, *Arms and Influence* (New Haven, CT: Yale University Press, 1966), 69ff.

9 This might have been the case had Japan been able to compel the United States to stop dropping atomic weapons on Japanese cities in August 1945—but, for obvious reasons, they could not and did not.

10 Lucas Kello, *The Virtual Weapon and International Order* (New Haven, CT: Yale University Press, 2017), 205–211.

11 NATO, "Brussels Summit Communique," June 14, 2021, https://www.nato.int/cps/en/natohq/news_185000.htm.

12 For an introduction to this concept, see Michael P. Fischerkeller and Richard J. Harknett, "Deterrence is Not a Credible Strategy for Cyberspace," *Orbis* 61, no. 3 (2017): 381–393.

13 Ellen Nakashima, "U.S. Cyber Command Operation Disrupted Internet Access of Russian Troll Factory on Day of 2018 Midterms," *Washington Post*, February 27, 2019, https://www.washingtonpost.com/world/national-security/us-cyber-command-operation-disrupted-internet-access-of-russian-troll-factory-on-day-of-2018-midterms/2019/02/26/1827fc9e-36d6-11e9-af5b-b51b7ff322e9_story.html.

14 David E. Sanger and Nicole Perlroth, "U.S. Escalates Online Attacks on Russia's Power Grid," *New York Times*, June 15, 2019, https://www.nytimes.com/2019/06/15/us/politics/trump-cyber-russia-grid.html.

and interfere with ransomware operations.[15] Recognizing that hostile cyberspace operations require an infrastructure to work efficiently, Persistent Engagement attacks this infrastructure to, at very least, raise the cost of hacking U.S. (and possibly allied) targets. But Persistent Engagement, at least in its current incarnation, focuses on hackers, not societies. It is not supposed to harm anyone aside from hackers.[16] There is also no indication that the United States intends to trade its right to undertake Persistent Engagement for any agreement on norms in cyberspace. In that sense, it was never meant to be a bargaining chip.[17] Its use is primarily tactical (notwithstanding that tactical measures can have strategic effects). In that sense, it is analogous to purely defensive measures meant to hinder offensive cyberspace operations by others. Last is the argument that Persistent Engagement can lead to "agreed competition,"[18] presuming it is used only against activities that the wielder of this capability would agree to not engage in (e.g., cyberattacks against critical infrastructure, political interference) and not used against activities that it would not agree to end (e.g., cyberespionage). But where is the evidence that such distinctions are being made? Furthermore, the similarity of cyberespionage and cyberattack in their early stages would suggest practical difficulties in telling them apart.

Standard deterrence is a one-step process: one side makes a threat, and the other side either does not start or does start and gets punished. In today's environment, where the base condition is cyberspace operations that the United States deems irresponsible state behavior, what happens is more that one side makes a threat and the other side either stops such operations or faces punishment. By contrast, obnoxious deterrence is a multi-step process. First, to modulate the other side's behavior, the one side (e.g., the United States) carries out its own cyberspace operations. The other side, understanding as much (that is, detecting, characterizing, and attributing these operations), enters negotiations to generate some ground rules in cyberspace, with the understanding that the price for curbing the one side's unwanted operations is to curb its own (in some cases, demands can be satisfied with actions such as arresting malware operators).

15 Ellen Nakashima and Dalton Bennett, "A Ransomware Gang Shut down after Cybercom Hijacked Its Site and It Discovered It Had Been Hacked," *Washington Post*, November 3, 2021, https://www.washingtonpost.com/national-security/cyber-command-revil-ransomware/2021/11/03/528e03e6-3517-11ec-9bc4-86107e7b0ab1_story.html.

16 If Persistent Engagement is also linked to implants in the Russian electric grid, then there is potential to create broader effects, but such effects are contingent, probably on the Russians doing something similar to the U.S. electric grid. As such, it bolsters standard deterrence against a specific type of cyberattack.

17 Indeed, Richard Harknett, one of the developers of the Persistent Engagement concept, writes that it can, "over time, lead to a normalization of cyberspace that is less free-for-all and potentially more stable." Richard Harknett, "United States Cyber Command's New Vision: What It Entails and Why It Matters," *Lawfare* [blog], March 28, 2018, https://www.lawfareblog.com/united-states-cyber-commands-new-vision-what-it-entails-and-why-it-matters. The key words are "over time," an indicator that this is not something to be traded away. See also Max Smeets, "US Cyber Strategy of Persistent Engagement and Defend Forward: Implications for the Alliance and Intelligence Collection," *Intelligence and National Security* 6, no. 3 (2020): 444–453; and Jason Healey, "The Implications of Persistent (and Permanent) Engagement in Cyberspace," *Journal of Cybersecurity*, 2019, 1–15.

18 Michael Fischerkeller and Richard Harknett, "Persistent Engagement, Agreed Competition, and Cyberspace Interaction Dynamics and Escalation," *Cyber Defense Review*, special edition, 2019, 267–287.

After agreement is reached, deterrence proceeds in a standard manner, the difference being that punishment for violations is the resumption of the behavior that motivated negotiations in the first place. Even if the temptation to cheat remains, at least both sides have a reasonable idea of what the consequences are: unilateral resumption leads to bilateral resumption. Conceivably, if one side figures that mutual resumption is in its interest, deterrence fails—but the same risk applies to standard deterrence as well.

3. WHAT OBNOXIOUS CYBERSPACE OPERATIONS FIT THE BILL?

However, determining what kind of obnoxious cyberspace operations will work to persuade the other side to stop its own obnoxious cyberspace operations is not trivial. Here are nine criteria that merit consideration and why they do. Meeting these criteria is not sufficient to guarantee success—the other side may just put up with the one side's obnoxious operations—but there are reasons to believe that each raises the odds that the other side will come to terms over their objectionable behavior.

The nine criteria below are explained from the perspective of a nation-state (referred to as the one side) that deters another nation-state that is a subject of this policy (the other side).

1. *The deterring activities should be cyberspace operations* rather than some other behavior the other side does not like. To understand why, consider the subsequent negotiations. At their most basic, they can enjoin many unwanted cyberspace operations: one (if one side's operations are like the other's), two (if one side's operations are sufficiently unlike the other's), or possibly more. But once something outside cyberspace is used for obnoxious deterrence, the negotiations must now cover two domains: where the one side operates (e.g., sanctions) and where the other side operates (cyberspace). The more compact the negotiations are, the fewer issues are raised, the less chance there is of distraction, and the easier they are to conclude. Even if the side initiating obnoxious cyberspace activities does not say why they started, the other side can guess that they have been undertaken to respond to the other side's cyberspace operations. If the one side's obnoxious operations are elsewhere (e.g., in space), the linkage will be far less obvious.
2. *The obnoxious cyberspace operations should work*. That is, they should succeed technically and noticeably. In most other domains, this goes without saying, but in cyberspace, it needs noting. Not all operations work, particularly against well-defended targets—and targets tend to be better-defended to the same degree that their owners cannot tolerate them being

damaged. It may not be obvious which class of targets are *ipso facto* hard before they are attacked. Attacking them in ways that alert the other side that something bad is taking place is apt to induce efforts to make them harder to attack, if only by closing the specific vulnerabilities that obnoxious cyberspace operations reveal.

3. *They should keep working*. If the target of obnoxious cyberspace operations can neutralize them through straightforward technical countermeasures (e.g., intrusion detection kits), then its incentive to negotiate them away is weak. As noted, this is a more serious consideration for cyberspace operations—which work to the extent that the target is vulnerable or otherwise unprepared—than it is for many military operations. But completely mitigating obnoxious cyberspace operations may itself be hard. Technical countermeasures may be expensive and complex; defenders worldwide collectively spend more than U.S. $100 billion on cyber defense[19] and still suffer attacks, notably involving ransomware. Furthermore, the ability to generate one novel attack suggests a heightened potential to carry out another novel attack—even if the original novel attack no longer works well. The fear of mass-casualty airline hijackings rose after 9/11, even though their odds of success declined in the face of later countermeasures (e.g., passengers charging the cockpit; later, hardened cockpit doors). Nullifying one cyberattack method is likely to heighten, not ease, the fear of undiscovered penetrations. Consider that Iran's hesitancy to install centrifuges after Stuxnet was discovered dented their build-up more than the number of centrifuges lost to Stuxnet directly.[20]
4. *The other side will not like it*. To wit, obnoxious cyberspace operations must not only succeed but also deliver politically salient discomfort. The latter depends on the other side's general resilience and its politics (i.e., who gets hurt and how that affects state decision-making). With cyberspace operations, the deterrer must also understand what capabilities of the other side would be substituted for those crippled by obnoxious cyberspace operations and, conversely, which operations may produce cascading and spillover effects (e.g., the closure of Colonial Pipeline's physical infrastructure after a ransomware attack that had no discernible effect on that infrastructure[21]). Put another way, the obnoxious behavior has to be of a quality or, failing that, quantity to be an effective impetus to negotiation.

19 Mary K. Pratt, "Cybersecurity Spending Trends for 2022: Investing in the Future," CSO, December 20, 2021, https://www.csoonline.com/article/3645091/cybersecurity-spending-trends-for-2022-investing-in-the-future.html.

20 David Albright, Paul Brannan, and Christina Walrond, "Stuxnet Malware and Natanz: Update of ISIS December 22, 2010 Report," Institute for Science and International Security, February 15, 2011, http://isis-online.org/isis-reports/detail/stuxnet-malware-and-natanz-update-of-isis-december-22-2010-reportsupa-href1/.

21 David E. Sanger, Clifford Krauss, and Nicole Perlroth, "Cyberattack Forces a Shutdown of a Top U.S. Pipeline," *New York Times*, May 8 and 13, 2021, https://www.nytimes.com/2021/05/08/us/politics/cyberattack-colonial-pipeline.html.

5. *The one side will get something from it*. The other side needs to be convinced that the deterrer is willing to carry out obnoxious cyber operations for a long time, perhaps indefinitely. Its belief that the deterrer is getting something from it would help make the case. Additionally, the more value the other side believes the deterrer is getting from its obnoxious cyberspace operations, the more it thinks it will have to offer to get the deterrer to quit. That noted, it is not obvious what Russia, for instance, is getting from much of its cyber mischief (e.g., its cyberattacks on the 2018 Winter Olympics). It would not be unprecedented for the other side to think that its adversaries are motivated solely by malevolence—but that motive may be good enough to sustain a long-term effort of obnoxious cyberspace activities.
6. *The deterrer will not get too much from it*. For an activity to be a bargaining chip, the actor must be willing to throw it into the pot in order to get something more valuable. There have been successful bargaining chips in U.S. strategic history, notably the Pershing II tactical nuclear missile (which was developed and deployed in the early 1980s, only to be eliminated when the Soviet Union signed the Intermediate-Range Nuclear Forces Treaty and eliminated its own SS-20 missiles). But the more investment has gone into a bargaining chip, the greater the pressure to keep it—both from the community that created it and those who could benefit from it. A politically strong leader may view the matter strategically and ignore these communities, but even strong leaders need to pick their battles. From the other side's perspective, the more the one side values the opportunities created by the relevant cyberspace operations (e.g., because it confounds the other side's censorship apparatus), the more that side may be tempted to cheat and carry out such operations even after it has agreed to forswear them. At some point, the other side may conclude that the odds of getting a credible and reasonable deal are too low to merit suing for relief.
7. *The deterrer can credibly negotiate their own actions away*. To wit, the behavior can be monitored,[22] and so, the other side can tell when the obnoxious activity stops, for instance, because the effects stop. At a minimum, that requires that cyberspace operations are definable with minimal ambiguity and that violations of any agreement are detectable or even measurable. By way of distinction, if cyberespionage is the obnoxious activity and the take, so to speak, stays in-house, the other side may not know the activity has stopped if hackers tread carefully. Other cyberspace operations (e.g., a cyberattack that stays quiet until called on) are meant to remain undiscovered; the same is true for data and algorithm corruption in small doses. Lastly, cyberspace operators should be counted on to stop when told to do so. This is not a problem in the United States, where command-and-control arrangements are solid and law enforcement is more than

22 See, for instance, John S. Davis II et al., *Stateless Attribution: Toward International Accountability in Cyberspace* (Santa Monica, CA: RAND, 2017), https://www.rand.org/pubs/research_reports/RR2081.html.

willing to go after private hackers. But countries that do not enforce their own cybersecurity laws or whose employees moonlight as hackers would have to work hard to be able to offer such promises credibly.

8. *The deterrer still comes out "ahead" if the other side starts doing likewise*. The other side is less likely to negotiate away its own activities if it believes that it can get the one side to back off by intensifying its own cyberspace activities (if the one side is doing the same thing the other side is doing) or doing likewise (if the one side is doing something different). The one side would have little basis for complaining—particularly if admitting, or at least not denying, such operations are key to its own negotiating strategy.
9. *There is no obvious escalation path forward for the other side*. It is usually in the one side's interest that the other side not escalate the conflict. Doing so raises costs; even if matters later de-escalate, the ability of the other side to dominate at higher levels of conflict—or at least prove a willingness to intensify the conflict and not yield—may doom any obnoxious deterrence strategy. Putatively, the odds of escalation reflect how invested the other side is in not backing down and the relative power of each side at higher levels of conflict. But an important secondary factor may be the nature of the one side's cyberspace operations: does it allow the other side to escalate without it being accused of, essentially, starting war anew? The more the other side can avoid such a narrative, the lower those barriers are to escalating (unless the other side wants its escalation to be obvious, so as to underline its power). Nevertheless, it is unclear what kind of obnoxious cyberspace operations would or would not suggest an obvious escalation path. Would one side's cyberattack on critical infrastructure lead to, say, a kinetic attack on the critical infrastructure of the other side, or would the latter be seen as war de novo? Would, say, the jamming of GPS in the one side's homeland (e.g., by using intermittent disposal emitters) be seen similarly? Would cyberattacks on military targets make escalation to the kinetic level similarly fraught for the other side? What about cyberattacks on the other side's domestic intelligence (e.g., secret police): does the other side have an obvious escalation path if the one side's domestic intelligence is clearly less critical to regime survival? This question can be reversed: if the one side's cyberattacks do not impress the other side, are there escalatory options that may be more persuasive but themselves do not feel like war de novo?

Most of the nine criteria inhibit an effective counter-strategy, one that would, variously, allow the other side to nullify the one side's cyberspace operations (that it works and that it keeps working), ignore them (that they will not like it), outlast them (that you get something from it), deprecate the value of negotiations (that you do not get too much from it), or force you to back off (that you can come out "ahead" even

if they do likewise and that they do not have a good escalation path). The other two—that it be a cyberspace operation and that you can credibly promise to stop—exist to facilitate negotiations. Many of these criteria, one should note, apply to after-the-fact retaliatory cyberspace operations. But again, with standard deterrence, the onus is on the one side to get the other side to quit. With obnoxious deterrence, the onus is placed on the other side to get the one side to quit. Likewise, issues of attribution (that the other side did what it was accused of) and characterization (that what the other side did was unacceptable practice) apply to standard deterrence. But with obnoxious deterrence, the difficulties of attribution and characterization apply to the other side as well. The other side cannot come to the table without, in effect, admitting that it has something to trade away: its own unwanted cyberspace operations.

Sanctions can more easily pass most of these criteria: they can work, they can keep on working (although the other side can reduce the pain over time by finding other markets or places to store assets), they can hurt (at least sometimes), the United States does not get too much value from levying them (but they cost the United States little), they can be credibly negotiated way (sort of; see below), the United States comes out "ahead" even if the other side does likewise, and there is no obvious escalation path upwards from sanctions. So they are relatively easy to design or select from. But they are not necessarily painful, are self-limited (e.g., you cannot stop buying something from someone twice), and are weak signals (being relatively painless to impose, they do not say how serious one is about the issue). They are, counterintuitively, harder to negotiate away: the United States has no intention of abjuring future sanctions—they are often applied as a punishment for a wide variety of sins, and, in practice, they linger well after the original event has faded into the past. Sanctions against Iraq that were imposed in the Saddam Hussein era did not end until 2010.[23] The use of sanctions puts pressure on the sanctioning country to make them work; they are deemed to have failed if the sanctioned country does not alter its behavior.[24]

Exactly what operations would fit these criteria would depend on who needs to be persuaded to stop something. It should be tailored to the other side's fears but not so deeply as to be seen as an existential challenge warranting an existential, hence escalatory, response. A candidate list of cyberspace operations may include doxing, challenges to information suppression (e.g., surmounting China's Great Firewall), cyberattacks that paralyze government-imposed processes required for transactions,[25] or those that would keep weapons systems from working correctly. In many cases,

23 "UN Lifts Sanctions against Iraq," BBC, December 15, 2020, https://www.bbc.com/news/world-middle-east-12004115.

24 See, for instance, Daniel Drezner, "The United States of Sanctions: The Use and Abuse of Economic Coercion," *Foreign Affairs*, September/October 2021, 142–154.

25 One such case was a cyberattack shut down Iranian gasoline stations because a government system was required to process a claim for subsidized gasoline. See Vivian Yee, "Iranian Motorists Hit with Cyberattack at Filling Stations," *New York Times*, October 26, 2021, https://www.nytimes.com/2021/10/26/world/middleeast/iran-gas-station-hack.html.

however, cessation of such cyberattacks would not erase all of its impacts (e.g., revealed information from doxing does not disappear when doxing does).

4. CYBERSPACE NARRATIVES

The most salient case against obnoxious deterrence is that using it interferes with the narrative that the United States wants to maintain about itself. In brief, the United States (together with its allies) favors a rules-based world order. Some state behaviors are considered responsible; other state behaviors are not. When countries carry out certain types of cyberspace operations (albeit a set that is still ill-defined), they break those rules. It is then up to responsible (and sufficiently powerful) countries to punish rule breakers, both to influence their behavior and to restore the moral order that dictates that those who are irresponsible do not benefit from their irresponsibility. But punishment so levied presumes the moral superiority of those who consider themselves responsible. That is an assumption that other states may and do resent and see as hypocritical.

The antithesis of a rules-based order is a power-based order, in which, as per the Melian Dialogue, the strong do as they will and the weak suffer what they must. Russia and China practice such realpolitik on a day-to-day basis. Despite paying lip service to a rule-based order, they challenge the rules set by the West. Such rules are deemed just another tool by which the powerful get their way.

Practicing obnoxious deterrence may require a narrative that countries must do things they would prefer no one do in the greater interest of establishing, one day, a true rules-based regime. This can be a hard argument to make. It is easier to argue that such actions legitimize bad behavior by committing it—thus making our pleas to stop look hypocritical. So there remains a tradeoff between narrative and realpolitik. This then raises the question whether hostile activity in cyberspace is sufficiently costly or risky that ending it justifies carving out a *temporary* exception to a rules-based order or whether a rules-based order is an ideal that increasing great-power rivalry makes a more distant dream.

The distinction between a rules-based and a power-based regime comes into play when one side negotiates away its obnoxious cyberspace operations. In a rules-based regime, the negotiations would focus on norms: the cyberspace operations carried out by the one side *against anyone* are off-limits in return for such cyberspace operations being off-limits by the other side against anyone. In a power-based regime, there is no necessary impetus to deem or declare certain actions to be irresponsible or otherwise forbidden; the one side promises to stop actions against the other side that it

dislikes and vice versa. But both are free to use them against unaligned third parties. In a bilateral world, the two (not doing it against the other side, and not doing it at all) are the same, but today's geopolitical environment has three cyber superpowers and many other highly competent players, including non-state actors. As the nuclear community is becoming painfully aware, given China's growing nuclear arsenal,[26] mutual forbearance among three parties is far trickier than among two. Here, at least, an agreement among each of two sides to holster some of its weapons against the other but not against third parties is plausible. Correspondingly, it is also plausible to view cyberspace negotiations as working within the context of bilateral relationships rather than being an exercise in multilateral norms-making.[27] Bilateral negotiations may be a gateway to multilateral norms—much as President Xi's promise to President Obama to halt economically motivated cyberespionage was followed by a broad G20 norm against such behavior.[28] But successful bilateral negotiations may remove some of the impetus to bring other countries into the agreement, particularly if most of the behavior the United States objects to comes from one country (much as China was responsible for the vast majority of economically motivated cyberespionage). Conversely, a series of bilateral negotiations may become progressively harder as each side presses for a deal as good as the last party negotiated ("why can't I have what you gave her?").

5. CONCLUSION

Obnoxious deterrence is symmetric, much as nuclear deterrence among superpowers is, and as judicial deterrence is not (the justice system penalizes criminals for misbehavior, but criminals are rarely in a position to reverse these roles). Many of the frustrations that occur with conventional deterrence—notably the difficulty of modifying the behavior of threat actors—would seem to be alleviated with a symmetric approach. If nothing else, the threat actors would have a countervailing interest in modifying U.S. behavior and a path, through negotiations, that would help them do so. And countries are more apt to follow rules that they have had a hand in crafting (even if done under pressure) as opposed to those imposed on them.

But obnoxious deterrence has its own difficulties. It is a multi-step process consisting of negotiations and enforcement (not just enforcement of unilateral red lines). Getting there requires mounting cyberspace operations, but only those that meet many criteria noted above. Both standard and obnoxious deterrence risk escalation, albeit at

26 Helene Cooper, "China Could Have 1,000 Nuclear Warheads by 2030, Pentagon Says," New York Times, November 3, 2021, https://www.nytimes.com/2021/11/03/us/politics/china-military-nuclear.html.

27 Cuba, which had virtually no presence in cyberspace, nevertheless managed to interfere with UN negotiations over cyberspace norms; see Ann Väljataga, "Back to Square One? The Fifth UN GGE Fails to Submit a Conclusive Report at the UN General Assembly," CCDCOE, https://ccdcoe.org/incyder-articles/back-to-square-one-the-fifth-un-gge-fails-to-submit-a-conclusive-report-at-the-un-general-assembly/.

28 Katie Bo Williams, "G20 Nations Reach Anti-Hacking Pledge," The Hill, November 17, 2015, https://thehill.com/policy/cybersecurity/260414-g20-nations-reach-anti-hacking-pledge.

different parts of their cycle. The narratives associated with standard and obnoxious deterrence are very different.

It is unclear that obnoxious deterrence would work in the sense of being more likely to modulate hostile behavior in cyberspace than to exacerbate it. But it is not as if conventional deterrence, as applied to sub-nuclear complaints, has worked any better. Indeed, one of the purposes of introducing the notion of obnoxious deterrence is to illuminate the weak prospects for deterrence in general.

A broader lesson may be in order. Since the Soviet Union disintegrated, the United States has enjoyed a period of asymmetric power vis-à-vis its rivals. This did not mean that it got everything it wanted, but it did suggest a correspondingly moral asymmetry. The United States could plausibly use its own behavior as the standard by which other countries' behavior could be judged, and, if found wanting, justify punishment. But the era of asymmetric power is giving way to an era of more symmetric power. As much as standard deterrence appeared to be a good fit for a period of asymmetric power, it may be a poor fit for a period of symmetric power. Correspondingly, obnoxious deterrence may be a bad idea whose time has come.

ACKNOWLEDGEMENTS

The author would like to acknowledge Nadia Kostyuk, Christopher Whyte, Jenny Jun, and other members of the Digital Issues Discussion Group for valuable feedback on a draft version of this paper. Their comments greatly improved its content.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

The Promise and Perils of Allied Offensive Cyber Operations

Erica D. Lonergan*
Assistant Professor
Army Cyber Institute at West Point
United States Military Academy
West Point, NY, United States
erica.lonergan@westpoint.edu

Mark Montgomery*
Senior Director
Center on Cyber and Technology Innovation
Foundation for Defense of Democracies
Washington, DC, United States
mark@cybersolarium.org

Abstract: NATO strategy and policy has increasingly focused on incorporating cyber operations to support deterrence, warfighting, and intelligence objectives. However, offensive cyber operations in particular have presented a delicate challenge for the alliance. As cyber threats to NATO members continue to grow, the alliance has begun to address how it could incorporate offensive cyber operations into its strategy and policy. However, there are significant hurdles to meaningful cooperation on offensive cyber operations, in contrast with the high levels of integration in other operational domains. Moreover, there is a critical gap in existing conceptualizations of the role of offensive cyber operations in NATO policy. Specifically, NATO cyber policy has focused on cyber operations in a warfighting context at the expense of considering cyber operations below the level of conflict. In this article, we explore the potential role for offensive cyber operations not only in wartime but also below the threshold of armed conflict. In doing so, we systematically explore a number of challenges at the political/strategic as well as the operational/tactical levels and provide policy recommendations for next steps for the alliance.

Keywords: *cyberspace, deterrence, NATO, offensive cyber operations*

* The views expressed in this article are personal and do not reflect the policy or position of any US government entity or organization.

1. INTRODUCTION

The North Atlantic Treaty Organization (NATO) is in the midst of conducting a comprehensive strategic initiative, "NATO 2030," to assess how to strengthen the alliance and shore up collective defense in an evolving geopolitical environment. While NATO 2030 is oriented toward the future strategic environment, cyber operations have already been testing the alliance's deterrence and defense strategy for well over a decade—especially with respect to adversarial cyber activity that takes place below the use-of-force threshold. In June 2021, the alliance convened in Brussels to affirm an agenda for NATO 2030.[1] Cybersecurity featured prominently in the Brussels communiqué, released during the summit, with the alliance emphasizing that cyber threats are "complex, destructive, coercive, and becoming ever more frequent."[2] Significantly, in Brussels the allies committed to a new comprehensive cyber defense policy. While the details of the policy remain opaque, the communiqué suggests some essential elements: that cyberspace represents a core aspect of the alliance's overall strategy of deterrence and defense; that the alliance is committed to "employ the full range of capabilities to actively deter, defend against, and counter the full spectrum of cyber threats"; and that a cyber attack could trigger the alliance's mutual defense obligations under Article 5 of the North Atlantic Treaty.[3] Indeed, NATO Secretary General Jens Stoltenberg recently emphasized that Article 5 applies to cyber attacks in a February 2022 press conference in the context of the Ukraine conflict, noting, "We have stated that cyber attacks can trigger Article 5."[4] However, NATO has not explicitly defined what would rise to the level of an armed attack in the cyber domain. Stoltenberg implied as much at the press conference, noting, "We have never gone into the position where we give a potential adversary the privilege of defining exactly when we trigger Article 5."[5]

One feature that stands out in the 2021 Brussels communiqué and the new cyber defense policy is the role of offensive cyber operations in NATO strategy. This is consistent with how the alliance's cyber policy has evolved in recent years, especially following the 2018 Brussels summit, where allies not only reaffirmed cyberspace as a domain of military operations but also publicly referenced efforts to integrate cyber effects operations voluntarily provided by allies into NATO military missions.[6] At a

1 "Leaders Agree NATO 2030 Agenda to Strengthen the Alliance," NATO, June 15, 2021, https://www.nato.int/cps/en/natohq/news_184998.htm.

2 "Brussels Summit Communiqué: Issued by the Heads of State and Government Participating in the Meeting of the North Atlantic Council in Brussels 14 June 2021," NATO, June 14, 2021, https://www.nato.int/cps/en/natohq/news_185000.htm.

3 "The North Atlantic Treaty," NATO, Washington, DC, April 4, 1949, https://www.nato.int/cps/en/natolive/official_texts_17120.htm.

4 "Press Conference by NATO Secretary General Jens Stoltenberg Following the Extraordinary Virtual Summit of NATO Heads of State and Government," NATO, February 25, 2022, https://www.nato.int/cps/en/natohq/opinions_192455.htm.

5 Ibid.

6 "Brussels Summit Declaration: Issued by the Heads of State and Government Participating in the Meeting of the North Atlantic Council in Brussels 11–12 July 2018," NATO, July 11, 2018, https://www.nato.int/cps/en/natohq/official_texts_156624.htm.

May 2019 speech in London, Stoltenberg emphasized that effective cyber deterrence demands a willingness to respond to attacks with the full range of capabilities and highlighted that the alliance is integrating offensive cyber capabilities into its military operations and missions.[7] This reflects the reality that cost-imposition is an integral part of any deterrence approach.[8]

Yet how exactly NATO should conceptualize the role of offensive cyber operations as part of its deterrence approach has raised important, unanswered questions. While the NATO alliance is anchored in traditional deterrence concepts—where the credible threat of (collective) retaliation to impose costs serves to dissuade an adversary from attacking any NATO ally—adversaries such as Russia exploit what is known as the "gray zone" short of war in a way that confounds and complicates traditional deterrence postures.[9] Moreover, these deterrence challenges are compounded by some of the unique aspects of offensive cyber operations, including intelligence-sharing requirements, planning processes and timelines, the nature of offensive cyber effects, and implications for sovereignty, that create potential impediments to a seamless integration of offensive cyber capabilities into existing NATO plans and force structures. Together, these factors illuminate important issues—not only at the strategic and political level but also at the operational and tactical level—about how the alliance could and should integrate offensive cyber operations into its strategy, planning, and operations. In this article, we argue that the alliance's primary focus on incorporating offensive cyber operations into conventional military planning is mismatched to the reality of the threat environment. In particular, we recommend that NATO should widen the aperture through which it considers the strategic value of offensive cyber capabilities, while systematically identifying and assessing the impediments the alliance may face in doing so. We first review the role of offensive cyber operations in deterrence strategies more broadly. Next, we discuss how NATO's approach to offensive cyber operations has evolved over time. We then explore the challenges of offensive cyber operations above and below the level of armed conflict, assessing gaps in existing NATO strategy and policy in this area. After that, we evaluate potential challenges and risks in incorporating a role for offensive cyber operations below the level of armed conflict. We conclude by offering a potential roadmap for how the NATO alliance should conceptualize offensive cyber operations going forward.

7 "Remarks by NATO Secretary General Jens Stoltenberg at the Cyber Defence Pledge Conference, London," North Atlantic Treaty Organization, May 23, 2019, https://www.nato.int/cps/en/natohq/opinions_166039.htm.

8 Joseph S. Nye, Jr., "Deterrence and Dissuasion in Cyberspace," *International Security* 41, no. 3 (Winter 2016/17): 44–71.

9 Janne Hakala and Jazlyn Melnychuk, "Russia's Strategy in Cyberspace," NATO Strategic Communications Centre of Excellence, June 2021; Lilly Pijnenburg Muller and Tim Stevens, "Upholding the NATO Cyber Pledge: Cyber Deterrence and Resilience; Dilemmas in NATO Defense and Security Policies," Norwegian Institute for International Affairs, Policy Brief 5 (2017): 1–4.

2. OFFENSIVE CYBER OPERATIONS IN DETERRENCE STRATEGIES

Deterrence strategies aim to shape the behavior of an adversary by influencing its perception of the costs, benefits, and risks of taking a particular action in order to dissuade it from doing so.[10] In practice, deterrence could come in different forms: punishment approaches threaten to impose significant costs on a target; denial approaches aim to make it more difficult for an adversary to carry outs its military strategy; entanglement uses the parties' interdependence to dissuade unwanted action; and norms leverage the risk of the target incurring reputational costs for defecting.[11] Successful deterrence—while difficult to observe, given that its result is the absence of an attack—occurs when the target assesses that the deterring state (or, in the case of NATO, alliance) has the capability and the political will to carry out the terms of its threat, and that the costs of disregarding the demand outweigh the prospective benefits.[12] As a result, successful deterrence in the context of an alliance is inherently more fraught than when conducted by a single state.[13]

With the emergence of cyberspace as a domain of strategic competition or conflict between states, scholars explored the extent to which deterrence logics developed for conventional and nuclear contests could be extended to this new area. Much of the early cyber deterrence literature was organized around drawing inferences about cyberspace based on the logic of deterrence in the nuclear age, which, by definition, was focused on decisive, strategic-level dynamics.[14] A core question researchers grappled with was the extent to which the threat of retaliation through the employment of offensive cyber power could be sufficiently credible and capable to prevent threat actors from engaging in undesirable behavior in cyberspace. In other words, this was focused on within-domain deterrence logics, with an emphasis on cost-imposition approaches. With nuclear deterrence as the reference point, researchers found that this form of cyber deterrence was highly problematic due to a range of factors: the challenge of rapid and accurate attribution; the limitation on information-sharing due

10 Glenn Snyder, *Deterrence and Defense: Toward a Theory of National Security* (Princeton, NJ: Princeton University Press, 1961), 9; Lawrence Freedman, Deterrence (Cambridge, UK: Polity, 2004), 26; Robert J. Art, "To What Ends Military Power?" *International Security* 4, no. 4 (1980): 3–35.

11 Nye, "Deterrence and Dissuasion in Cyberspace"; John J. Mearsheimer, *Conventional Deterrence* (Ithaca, NY: Cornell University Press, 1985): 14–15.

12 John J. Mearsheimer, "Nuclear Weapons and Deterrence in Europe," *International Security* 9, no. 3 (1984): 21.

13 One could argue that it could be sufficient for at least one party to the alliance to have these capabilities. However, this poses credibility challenges for the alliance as a whole, particularly in terms of the adversary's perception. This is one reason NATO has prioritized interoperability for conventional warfighting.

14 Martin Libicki, *Cyberdeterrence and Cyberwar* (Santa Monica, CA: RAND Corporation, 2009); Richard L. Kugler, "Deterrence of Cyber Attacks," in *Cyberpower and National Security*, ed. Franklin D. Kramer, Stuart H. Starr, and Larry K. Wentz (Washington, DC: National Defense University Press, 2009); Emily O. Goldman and John Arquilla, *Cyber Analogies* (Monterey, CA: Naval Postgraduate School, 2014); Ben Buchanan, *The Cybersecurity Dilemma: Hacking, Trust, and Fear Between Nations* (New York: Oxford University Press, 2016).

to secrecy about cyber techniques; the absence of shared international understandings about acceptable behavior; the "borderless" nature of cyberspace; the speed of attack; low barriers to entry; the diversity of cyber actors; and limitations on obtaining violent effects through the use of offensive cyber power.[15] This led some to reject cyber deterrence as a feasible strategic approach.[16]

However, recent research has reframed the cyber deterrence debate to move away from nuclear-era frameworks to explore the feasibility of cross-domain cyber deterrence, as well as how offensive cyber operations could function as part of deterrence by denial (rather than punishment) strategies, particularly below the level of armed conflict. In this view, offensive cyber operations are not conceived of as a mechanism for punishment strategies to be held in reserve and threatened as a form of retaliation. Instead, this literature evaluates how counter-cyber operations that disrupt, deny, degrade, and deceive adversary offensive cyber capabilities and attack infrastructure, as well as the military strategy and organization that enables offensive campaigns, could achieve deterrence effects short of armed conflict.[17] Building on this literature, we explore how this type of deterrence approach could work in the context of NATO strategy and policy in light of NATO's move to incorporate offensive cyber operations into its approach.

3. A GROWING ROLE FOR OFFENSIVE CYBER OPERATIONS IN NATO STRATEGY AND POLICY

Russia's 2007 cyber attacks against Estonia were a galvanizing event for the NATO alliance, prompting NATO to adopt a policy on cyber defense at the Bucharest summit in 2008 and to establish the NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) in Tallinn.[18] Since then, the alliance has increasingly prioritized cybersecurity issues at successive ministerial-level meetings and summits, created

15 Lucas Kello, *The Virtual Weapon and International Order* (New Haven, CT: Yale University Press, 2017); John Arquilla and David Ronfeldt, "Cyberwar Is Coming," *Comparative Strategy* 12, no. 2 (1993): 141–165. For competing views, see Erik Gartzke, "The Myth of Cyberwar: Bringing War in Cyberspace Back Down to Earth," *International Security* 38, no. 2 (2013): 41–73; Thomas Rid, "Cyber War Will Not Take Place," *Journal of Strategic Studies* 35, no. 1 (2012): 5–32; Brandon Valeriano, Benjamin M. Jensen, and Ryan C. Maness, *Cyber Strategy: The Evolving Character of Power and Coercion* (New York: Oxford University Press, 2018).

16 Michael Fischerkeller and Richard Harknett, "Deterrence is Not a Credible Strategy," *Orbis* 61, no. 3 (2017); Michael P. Fischerkeller, "Persistent Engagement and Tacit Bargaining: A Strategic Framework for Norms Development in Cyberspace's Agreed Competition," *Institute for Defense Analysis*, November 2018.

17 Jacquelyn G. Schneider, "Deterrence in and through Cyberspace," in *Cross-Domain Deterrence: Strategy in an Era of Complexity*, ed. Erik Gartzke and Jon R. Lindsay (New York: Oxford University Press, 2019); Jon Lindsay et al., "Cybersecurity and Cross-Domain Deterrence: The Consequences of Complexity," *Journal of Cybersecurity* 1, no. 1 (2015): 53–67; Erica D. Borghard and Shawn W. Lonergan, "Deterrence by Denial in Cyberspace," *Journal of Strategic Studies* (2021).

18 Joshua Davis, "Hackers Take Down the Most Wired Country in Europe," *Wired*, August 21, 2007; "Bucharest Summit Declaration: Issued by the Heads of State and Government in Participating in the Meeting of the North Atlantic Council in Bucharest on 3 April 2008," North Atlantic Treaty Organization, April 3, 2008.

new organizations and centers to address cyber threats, and incorporated cyber issues into NATO policy. These efforts have largely focused on the defensive aspects of cybersecurity, with the alliance maintaining that its "main focus in cyber defense is to protect its own networks (including operations and missions) and enhance resilience."[19] For instance, NATO established a Computer Incident Response Capability (NCIRC) to defend NATO networks and systems and enable swift information-sharing, along with Rapid Reaction Teams—defensive teams that can be quickly deployed to assist NATO and allied networks.[20] It also created the NATO Communications and Information Agency (NCIA), combining three different organizations, to defend NATO's communications systems against cyber attacks. At the 2014 summit in Wales, NATO members endorsed the Enhanced Cyber Defence Policy, which affirmed that cyber defense is an inherent aspect of collective defense and that the alliance would incorporate cyber defense into its planning and operations.[21] In 2016, alliance members agreed to a Cyber Defence Pledge in which they committed to improving their cyber defenses and emphasized a range of defense initiatives, including training, education, exercises, and information-sharing.[22]

At the same time, since 2016, the offensive aspects of cyber operations have become more prominent in NATO policy.[23] At the 2016 Warsaw summit, the alliance affirmed the agreement reached in June of that year among NATO defense ministers that cyberspace should be recognized as a military domain.[24] Furthermore, the NATO Cyber Operations Center, which was created at the 2018 Brussels summit and reached initial operational capability in the spring of 2021, serves to coordinate requests for offensive cyber effects from member states through a process termed "Sovereign Cyber Effects Provided Voluntarily by Allies."[25] After the 2018 Brussels summit, it was made public that five states—the United States, United Kingdom, Denmark, the Netherlands, and Estonia—were contributing cyber forces to NATO, but it is not clear what their purpose is.[26] Moreover, it was acknowledged that "offensive cyber effects are not yet part of the [NATO] mission planning process."[27]

19 "Cyber Defence," NATO, July 2, 2021, https://www.nato.int/cps/en/natohq/topics_78170.htm.

20 "NATO Rapid Reaction Team to Fight Cyber Attack," NATO, March 13, 2012, https://www.nato.int/cps/en/natolive/news_85161.htm.

21 "Wales Summit Declaration: Issues by the Heads of State and Government Participating in the Meeting of the North Atlantic Council in Wales," NATO, September 5, 2014, https://www.nato.int/cps/en/natohq/official_texts_112964.htm.

22 "Cyber Defence Pledge," NATO, July 8, 2016, https://www.nato.int/cps/en/natohq/official_texts_133177.htm.

23 James A. Lewis, "The Role of Offensive Cyber Operations in NATO's Collective Defence," Tallinn Paper No. 8 (2015). See also Don Lewis, "What is NATO Really Doing in Cyberspace?" *War on the Rocks*, February 4, 2019.

24 Lillian Ablon, Anika Binnendijk, Quentin E. Hodgson, Bilyana Lilly, Sasha Romanosky, David Senty, and Julia A. Thompson, "Operationalizing Cyberspace as a Military Domain: Lessons for NATO," RAND Corporation, July 2019.

25 Jeppe T. Jacobsen, "Cyber Offense in NATO: Challenges and Opportunities," *International Affairs* 97, no. 3 (May 2021): 703–704; "Cyber Defence," NATO.

26 "New Conference by Secretary Mattis at NATO Headquarters, Brussels, Belgium," US Department of Defense, October 4, 2018.

27 Ion A. Iftimie, "NATO's Needed Offensive Cyber Capabilities," NATO Defense College Policy Brief 10 (May 2020), 2.

4. CYBER DETERRENCE—AT WHAT THRESHOLD?

NATO's shift to incorporating offensive cyber operations into existing strategy and policy has focused on integrating offensive effects into conventional military plans and operations to be employed in the context of a conflict. This is meant to enhance NATO's deterrence posture (given that cyber activity is an inevitable part of conflict, signaling a willingness to use offensive cyber capabilities alongside other military capabilities, as well as demonstrating cyber capabilities, makes deterrence more credible), as well as NATO's collective defense capabilities against likely adversary cyber attacks if deterrence fails and conflict takes place. Echoing the development of NATO policy, experts have suggested that NATO develop a framework for incorporating offensive cyber operations into doctrine and planning. Yet this is similarly conceptualized as being part of warfighting, such as using offensive cyber operations to "disrupt an adversary's communications, logistics, and sensors."[28] Likewise, experts have explored how to synchronize cyber and kinetic effects on the battlefield as part of NATO's multi-domain operations warfighting concept.[29] NATO's increasing willingness to discern feasible approaches to incorporating offensive cyber operations into policy is a positive development. However, focusing solely on scenarios for the employment of cyber power during conflict—either to enhance or complement conventional campaigns—is mismatched to the cyber threat environment NATO confronts, to the reality of planning and conducting cyber operations, and to how a number of NATO members conduct offensive cyber operations (which has implications for the alliance as a whole).

First, the prevailing threat NATO allies currently face in the cyber domain does not stem from high-end, decisive cyber attacks in the midst of a broader military campaign. Rather, cyber threats manifest as a range of gray-zone tactics that in themselves do not rise to a level of use of force or armed attack but nevertheless have corrosive and, in some instances, strategic effects against the alliance.[30] Indeed, the language in the 2021 Brussels communiqué reflects this reality—that the accumulation of individual malicious incidents (such as cyber penetrations of critical infrastructure, large-scale cyber theft of industrial and economic information, and interference in democratic institutions and processes) could be sufficient to trigger the invocation of collective defense. In July 2021, NATO issued an official statement condemning a range of malicious cyber behavior, including the Microsoft Exchange hack (which the alliance publicly attributed to China), as well as ransomware attacks, but did not

28 Franklin D. Kramer, Robert J. Butler, and Catherine Lotrionte, "Cyber, Extended Deterrence, and NATO," Atlantic Council, Issue Brief (May 2016), 6; Lewis, "The Role of Offensive Cyber Operations," 10; Matthijs Veenendaal, Kadri Kaska, and Pascal Brangetto, "Is NATO Ready to Cross the Rubicon on Cyber Defence?" NATO CCDCOE Cyber Policy Brief, June 2016.

29 Franz-Stefan Gady and Alexander Stronell, "Cyber Capabilities and Multi-Domain Operations in Future High-Intensity Warfare in 2030," in *Cyber Threats and NATO 2030: Horizon Scanning and Analysis*, eds. Amy Ertan et al. (NATO CCDCOE, 2020): 151–176.

30 Ertan et al., *Cyber Threats*.

claim that any of these activities rose to the level of armed attack.[31] Cyberspace is only one manifestation of the broader gray-zone challenge the NATO alliance faces. One example, though not against a NATO member, is Russia's 2014 invasion of Ukraine and annexation of Crimea through the use of plausibly deniable proxy forces. However, this illustrates how NATO could soon be grappling less with the risk of a decisive armed attack on a member state than with the employment of hybrid and gray-zone tactics short of war to undermine and subvert the alliance in a way that avoids triggering the mutual defense clause of the NATO treaty.[32] Therefore, while NATO's emerging approach to offensive cyber operations is oriented toward one type of strategic challenge—how to integrate offensive cyber operations into conventional military planning in the event of an attack on a NATO member—it appears to leave unaddressed the day-to-day reality of the threat environment, in which threat actors (most importantly, Russia) seek to subvert and undermine the security and defense of the alliance through offensive cyber means short of war. NATO's current approach of incorporating offensive cyber operations into its conventional planning to enhance its deterrence posture for outright attacks is not applicable to malicious behavior below that threshold.[33]

A second mismatch is between NATO's concept for offensive cyber operations as part of conventional warfighting and the reality of the requirements for planning and conducting them. Specifically, offensive cyber operations demand a significant prior investment. This involves not only planning and preparation to be able to cause effects against a target during wartime (which is also the case for conventional military operations) but also cyber intelligence and exploitation operations to identify vulnerabilities, gain access, develop exploits, and continuously hold targets at risk until the desired time of employment.[34] As Austin Long notes, "the intelligence requirements for cyber options are immense, as the delivery mechanism is entirely dependent on intelligence collection."[35] Beyond the level of trust that would be required to bring offensive cyber operations into NATO, given the central role of intelligence, another important implication of this is that there is not a clear demarcation between cyber operations below and above the threshold of armed conflict. States necessarily have to breach adversary networks and systems and maintain access to critical targets before war takes place—even if NATO were to only employ offensive effects during wartime. This is particularly true when developing offensive cyber campaigns against a more sophisticated adversary like Russia, which is likely to be better defended

31 "Statement by the North Atlantic Council in Solidarity with Those Affected by Recent Malicious Cyber Activities Including the Microsoft Exchange Server Compromise," NATO, July 19, 2021, https://www.nato.int/cps/en/natohq/news_185863.htm?selectedLocale=en.

32 Maria Mälksoo, "Countering Hybrid Warfare as Ontological Security Management: The Emerging Practices of the EU and NATO," *European Security* 27, no. 3 (2018): 385–386.

33 Jacobsen, "Cyber Offense in NATO," 714–716.

34 Herbert Lin, "Offensive Cyber Operations and the Use of Force," *Journal of National Security Law and Policy* 4, no. 9 (2010): 63–86; Erica D. Borghard and Shawn W. Lonergan, "The Logic of Coercion in Cyberspace," *Security Studies* 26, no. 3 (2017): 452–481.

35 Austin G. Long, "A Cyber SIOP? Operational Considerations for Strategic Offensive Cyber Planning," *Journal of Cybersecurity* 3, no. 1 (2017): 22.

(and therefore present a greater number of hardened targets that require close access, custom exploits, and so on).[36]

Similarly, there is not always a clear delineation in practice between offensive and defensive action in cyberspace. Computer network exploitation is a necessary prior condition to attack and, as a result, it can be difficult to distinguish between cyber operations for intelligence collection and those meant to obtain effects. In addition to that, many defensive actions blur the lines between offense and defense.[37] Static defenses (such as firewalls, intrusion detection, access management, and other measures) are not sufficient to defend against nation-state adversaries, such as Russia, that have demonstrated the commitment and the capability to carry out malicious cyber campaigns against NATO members. Active defense and deception (such as threat hunting, baiting, decoys, and honeypots), which exist at an ambiguous nexus between offense and defense, are therefore an important part of cyber defense.[38] Therefore, even if NATO policy seeks to draw a clear distinction between offensive cyber operations during conflict and defensive operations in routine competition, it will need to consider how to define a role for some forms of defensive activities that are more proactive in nature (and may be perceived as offensive).

Finally, NATO's focus on offensive cyber operations above the level of armed conflict is in tension with how several NATO allies are already operating in the cyber domain. A growing number of NATO members have publicly expressed an interest in developing offensive cyber capabilities, but there is significant variation in offensive capabilities and the level of organizational maturity across the alliance.[39] Some NATO members began to develop strategies and organizations for military cyber organizations in 2008, and most are now either launching or expanding their offensive cyber programs.[40] From a policy perspective, some states have been more public about conducting offensive cyber operations. For instance, in the United States, cyber policy underwent a significant shift in 2018, when the Department of Defense and US Cyber Command issued updated strategy and policy documents that articulated a role for the military in conducting offensive cyber operations below the level of armed conflict outside of US-controlled cyberspace.[41] As part of its "defend forward" and "persistent engagement" approach, the United States has reportedly conducted a number of out-of-network cyber operations, such as the 2018 operation to disrupt the ability of the Internet Research Agency—a Russian troll farm—to interfere in the midterm elections, dropping "cyber bombs" on the so-called Islamic State to disrupt

36 Jacobsen, "Cyber Offense in NATO," 705, 709. See also Lewis, "The Role of Offensive Cyber Operations," 4–5.

37 Buchanan, *Cybersecurity Dilemma*.

38 Jacobsen, "Cyber Offense in NATO," 711; Erik Gartzke and Jon R. Lindsay, "Weaving Tangled Webs: Offense, Defense, and Deception in Cyberspace," *Security Studies* 24, no. 2 (2015): 316–348.

39 Max Smeets, "NATO Members' Organizational Path Toward Conducting Offensive Cyber Operations: A Framework for Analysis," 2019 11th International Conference on Cyber Conflict.

40 Smeets, "NATO Members' Organizational Path," 3, 7.

41 "Summary: Department of Defense Cyber Strategy 2018," US Department of Defense, https://media.defense.gov/2018/Sep/18/2002041658/-1/-1/1/CYBER_STRATEGY_SUMMARY_FINAL.PDF.

their communications and command and control, and more recently, disrupting ransomware groups targeting critical infrastructure.[42] The United States has also worked bilaterally with other NATO allies, such as Estonia[43] and Montenegro,[44] to conduct "hunt forward" cyber operations on allied and partner networks to uncover and disrupt malicious cyber activity. But the United States is not alone among NATO allies in being more transparent about offensive cyber operations. In 2020, the United Kingdom announced a significant investment in its National Cyber Force, which is its organizational arm for offensive cyber operations.[45] Moreover, UK leaders have become increasingly public in discussing Britain's role in offensive cyber operations.[46] Other NATO members, such as the Netherlands, have also publicly alluded to conducting offensive cyber operations.[47]

5. ADDRESSING CHALLENGES AND MANAGING RISKS

Given these mismatches, some experts have begun to call for a more flexible approach to cyber deterrence. Ion Iftimie, for example, argues that NATO allies should agree to a set of graduated cyber response options that provide more flexibility for the alliance to address cyber threats while managing potential escalation risks.[48] As part of this, Iftimie calls for NATO to "engage in active military measures to deny, degrade, disrupt, deceive, or destroy an adversary's offensive cyber capabilities," which he terms "cyber A2/AD capabilities."[49] While a more flexible approach would be better tailored to the nature of NATO's threat environment, how the alliance would actually implement this, to include potentially leveraging offensive cyber power short of war, remains underspecified. Therefore, systematically evaluating potential issues and impediments is important as the alliance assesses how to operationalize a role for offensive cyber power in practice. We organize these into three categories: strategic/political level issues, operational/tactical issues, and risks.

42 Erica D. Borghard, "What a U.S. Operation against Russian Trolls Predicts about Escalation in Cyberspace," *War on the Rocks*, March 22, 2019; David E. Sanger, "U.S. Cyberattacks Target ISIS in a New Line of Combat," *New York Time*s, April 24, 2016; Sean Lyngaas, "US Military's Hacking Unit Publicly Acknowledges Taking Offensive Action to Disrupt Ransomware Operations," CNN.com, December 5, 2021.

43 "Hunt Forward Estonia: Estonia, US Strengthen Partnership in Cyber Domain with Joint Operation," US Cyber Command, December 3, 2020, https://www.cybercom.mil/Media/News/Article/2433245/hunt-forward-estonia-estonia-us-strengthen-partnership-in-cyber-domain-with-joi/.

44 Julian Barnes, "U.S. Cyber Command Expands Operations to Hunt Hackers From Russia, Iran and China," *New York Times*, November 2, 2020.

45 Joe Devanny and Tim Stevens, "What Will Britain's New Cyber Force Actually Do?" *War on the Rocks*, May 26, 2021.

46 "UK and US Join Forces to Strike Back in Cyber-Space," BBC.com, November 18, 2021; "National Cyber Strategy 2022," United Kingdom, December 15, 2021, https://www.gov.uk/government/publications/national-cyber-strategy-2022/national-cyber-security-strategy-2022.

47 Max Smeets, "The Netherlands Just Revealed Its Cybercapacity. So What Does That Mean?" *Washington Post*, February 8, 2018.

48 Iftimie, "NATO's Needed Offensive Cyber Capabilities," 3.

49 Ibid., 4. Advocates sometimes sidestep some significant impediments. See Kramer et al., "Cyber, Extended Deterrence, and NATO," 7–10.

A core challenge at the political level is the reality that there is a lack of consensus among NATO allies about the appropriate application of offensive cyber power—especially below the level of armed conflict. While it is inevitable that allies will have varying approaches on matters of strategy and policy, directly addressing these differences is important for NATO's strategic and political cohesion. This is especially true for cyber operations because the implementation of offensive cyber operations by a given state may mean that it is maneuvering through or operating on networks controlled by its allies, potentially giving rise to friction between them.[50] Currently, there is "no agreement among NATO countries… concerning the procedures and limits of an offensive action within the cyber domain, particularly on access to systems and networks located in another allied country."[51] This is due, in part, to differences among allies about how to define sovereignty in cyberspace.[52] Therefore, allies may disagree on whether certain types of offensive cyber operations represent a violation of international law. There are also likely to be differences in how allies define the conditions under which offensive cyber operations would rise to the level of armed attack.[53] As a result, the announcement of a more proactive posture by the United States in its new defend-forward concept generated some concerns among its allies about appropriate parameters and conditions for consultation, notification, implications for norms, and how to manage potential inadvertent escalation.[54] Max Smeets, for example, calls for NATO allies to sign a memorandum of understanding regarding offensive cyber operations to clarify these issues.[55]

Even if NATO allies can reach shared understandings about whether, and how, offensive cyber operations could be employed short of war, and how to define appropriate boundaries and scope of these operations, there are several challenges at the tactical and operational level that should be considered when implementing this policy. For example, intelligence is an essential aspect of offensive cyber operations, but this complicates how offensive cyber could be incorporated into NATO doctrine and planning.[56] For one thing, there is likely to be tension between intelligence collection requirements, which are defined at the national level, and decisions about military operations, which in existing NATO structures would take place at the alliance level.[57] Moreover, beyond the scope of the Five Eyes partnership, NATO allies are likely to be quite circumspect about sharing intelligence that would enable

50 Max Smeets, "U.S. Cyber Strategy of Persistent Engagement and Defend Forward: Implications for the Alliance and Intelligence Collection," *Intelligence and National Security* 35, no. 3 (2020).
51 Alessandro Marrone and Ester Sabatino, "Cyber Defence in NATO Countries: Comparing Models," *Instituto Affari Internazionali* 21, no. 5 (February 2021): 13.
52 Michael N. Schmitt and Liis Vihul, "Respect for Sovereignty in Cyberspace," *Texas Law Review* 95, no. 7.
53 Michael N. Schmitt, ed., *Tallinn Manual on the International Law Applicable to Cyber Warfare* (Cambridge: Cambridge University Press, 2013); *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge: Cambridge University Press, 2017).
54 Joshua Rovner, "More Aggressive and Less Ambitious: Cyber Command's Evolving Approach," *War on the Rocks*, September 14, 2020.
55 Max Smeets, "NATO Allies Need to Come to Terms with Offensive Cyber Operations," *Lawfare*, October 14, 2019.
56 Lewis, "The Role of Offensive Cyber Operations," 7.
57 Ibid., 9.

effective offensive operations. This is due not only to general concerns about sharing, for instance, information gleaned from signals intelligence collection—among the most sensitive intelligence collection methods—but also to the reality that allies may be using the same exploits that would enable offensive action to conduct espionage, including potentially on an ally. In May 2021, for example, it was reported that the United States worked with the Danish government to spy on senior leaders in Germany, Sweden, Norway, and France.[58] While this particular example is about espionage rather than offensive cyber operations, it reveals the delicate challenges of intelligence-sharing in a context where offensive cyber operations depend heavily on intelligence collection. On a related note, deconflicting intelligence and military priorities—deciding the conditions under which to prioritize intelligence collection over military effects—is often contested within a particular state and would be even more challenging to adjudicate across an alliance.[59]

Finally, there is an obvious risk that moving toward a more offensive posture in cyberspace below the level of warfare could increase the likelihood of escalation and undermine NATO's broader deterrence approach. Jeppe Jacobsen, for instance, argues that there is a significant risk of escalation if NATO decides to serve a coordinating function for allied offensive cyber operations below the use of force threshold against Russia. Specifically, he postulates that "the likelihood increases that Russia misinterprets these effects as escalatory and acts accordingly."[60] While escalation concerns should not be ignored, there is a burgeoning body of academic research that has found little empirical support for the contention that cyber operations lead to escalation—especially to kinetic military conflict.[61] Rather than reject the feasibility of offensive cyber operations outright due to fears of escalation, the alliance should consider how to strengthen existing confidence-building measures, particularly with Russia, to enable more effective communication and transparency about cyber operations to mitigate the risks of unintended effects.[62]

58 "U.S. Spied on Merkel and Other Europeans through Danish Cables: Broadcaster DR," Reuters, May 31, 2021.

59 Jacobsen, "Cyber Offense in NATO," 712–713.

60 Ibid., 719.

61 Erica D. Borghard and Shawn W. Lonergan, "Cyber Operations as Imperfect Tools of Escalation"; Sarah Kreps and Jacquelyn Schneider, "Escalation Firebreaks in the Cyber, Conventional, and Nuclear Domains: Moving Beyond Effects-Based Logics," *Journal of Cybersecurity* 5, no. 1 (2019); Benjamin Jensen and Brandon Valeriano, "What Do We Know About Cyber Escalation? Observations from Simulations and Surveys," Atlantic Council Issue Brief (November 2019).

62 Erica D. Borghard and Shawn W. Lonergan, "Confidence Building Measures for the Cyber Domain," *Strategic Studies Quarterly* 12, no. 3 (Fall 2018): 10–49.

6. NEXT STEPS

The evolution of NATO policy over the past several years toward a recognition that there is a role for offensive cyber operations in its deterrence and collective defense strategy is a positive development. However, the focus on incorporating cyber operations into military planning for conflict does not account for the reality of the threat environment NATO allies currently face, which stems from malicious cyber behavior short of war. Therefore, the alliance should also deliberately—but carefully—consider a potential role for offensive cyber operations below the level of armed conflict. Frank conversations that address and aim to reconcile potentially competing perspectives among allies will be essential for the effectiveness of the alliance's cyber strategy. These discussions should be an important component of the NATO 2030 review.

One question raised by our analysis is whether the benefits of integrating offensive cyber operations within the NATO alliance are outweighed by the significant interoperability challenges and, as a result, whether continuing with a smaller, coalition-based approach is optimal. While interoperability issues should not be dismissed, for cyber operations to be a credible component of NATO's deterrence strategy, an ad hoc effort below the alliance level will likely be insufficient. That said, given the interoperability issues, the alliance should identify and explore solutions to potential gaps and challenges by incorporating offensive cyber operations below the level of armed conflict into existing NATO simulations and exercises across strategic, operational, and tactical levels.[63] Concerning the coordination of offensive cyber operations and roles and responsibilities, a number of questions remain unanswered. For instance, how could allies improve intelligence-sharing in support of more rapid attribution that would enable a member state or the alliance as a whole to respond to adversary cyber activity? From an offensive cyber perspective, could one ally share a tool with another ally that may have the requisite access to a target, or assign a trained operator to leverage that infrastructure if the tool is highly tailored and requires specific training? What are the conditions under which allies could share common toolkits or access to critical targets? Alternatively, what are the conditions under which allies should consider dividing responsibilities for cyber campaign planning and developing exquisite access and capabilities against hard targets in, for example, Russia? For those time-, resource- and personnel-intensive campaigns against a sophisticated adversary, sharing this burden across several states could increase the aggregate level of cyber power that can be brought to bear against it. Furthermore, if some allies are responsible for offensive cyber operations against certain targets, what are the information-sharing and notification requirements?

63 Ablon et al., "Operationalizing Cyberspace as a Military Domain," 12–15.

There are significant challenges associated with integrating cyber operations that occur in wartime, let alone when they are below the threshold of armed attack. Setting up a process to adjudicate political differences, address intelligence-sharing issues, and grapple with operational challenges will be difficult. However, there is a clear strategic imperative for NATO to do so—one that has only been underscored by Russia's recent cyberactivity in the context of its invasion of Ukraine. Moreover, addressing these difficult questions and building institutional processes for operating in cyberspace toward shared objectives short of war could have positive spillover effects for discerning how to improve interoperability above that threshold if deterrence fails.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

'Releasing the Hounds?'[1] Disruption of the Ransomware Ecosystem Through Offensive Cyber Operations

Michael Bátrla
Researcher
Faculty of Law
Masaryk University
Brno, Czech Republic
michael.batrla@law.muni.cz

Jakub Harašta
Assistant Professor
Faculty of Law
Masaryk University
Brno, Czech Republic
jakub.harasta@law.muni.cz

Abstract: Ransomware groups represent a significant cyber threat to Western states. Most high-end ransomware actors reside in territorial safe-haven jurisdictions and prove to be resistant to traditional law enforcement activities. This has prompted public sector and cybersecurity industry leaders to perceive ransomware as a national security threat requiring a whole-of-government approach, including cyber operations. In this paper, we investigate whether cyber operations or the threat of cyber operations influence the ransomware ecosystem. Subsequently, we assess the vectors of influence and characteristics of past operations that have disrupted the ecosystem. We describe the specifics of the ransomware-as-a-service system and provide three case studies (DarkSide/BlackMatter, REvil, Conti) highly representative of the current ecosystem and the effect cyber operations have on it.

Additionally, we present initial observations about the influence of cyber operations on the system, including best practices from cyber operations against non-state groups. We conclude that even professional, highly skilled, and top-performing ransomware groups can be disrupted through cyber operations. In fact, cyber operations can even

1 'Releasing the hounds' is a term for offensive cyber operations aimed at disrupting global ransomware gangs, especially those conducted by militaries or intelligence agencies. First use is found in Patrick Gray and Adam Boileau, 'Feature Podcast: Releasing the Hounds with Bobby Chesney', Risky Business, 28 May 2020, https://risky.biz/HF6/.

bypass some limits imposed on law enforcement operations. Even when ransomware groups rebrand or resurface after a hiatus, we suggest their infrastructure (both technical, human, and reputational) will still suffer mid- to long-term disruption. Although cyber operations are unlikely to be a silver bullet, they are an essential tool in the whole-of-government and multinational efforts and may even grow in importance in the next several years.

Keywords: *ransomware-as-a-service, cybercrime, offensive cyber operations, cyber incident, cryptocurrency*

1. INTRODUCTION

Ransomware groups represent a significant cyber threat to Western states' critical and non-critical infrastructure, as will be discussed in detail below. In 2020 and 2021, major incidents and campaigns raised awareness of the problem of ransomware and the growing professionalization of cybercriminals. The Colonial Pipeline attack and attacks on hospitals in the USA, Australia, Ireland, and the Czech Republic received the most coverage but represented just the tip of the iceberg. Healthcare services amidst the COVID-19 pandemic, educational institutions, international organizations, and national and local governments are all seeing increased ransomware activity. The cumulative consequences of such incidents are unprecedented for cybercrime.

In this paper, we mainly focus on high-end ransomware groups and ransomware-as-a-service programs. As will be shown below, these seem to cause most incidents across key public services and companies, imposing ever-growing costs. While most of the academic literature concentrates on the technical side and detection and prevention of attacks, we take a broader – more strategically oriented – approach. Most ransomware actors reside in territorial safe-haven jurisdictions (also known as bulletproof jurisdictions), which tend to be resistant to requests for judicial or law enforcement cooperation, being unlikely to prosecute crimes unless directed against their own interests, and effective law enforcement action is often hampered.

Consequently, both public sector and cybersecurity industry leaders describe ransomware as a national security threat that should be handled with a whole-of-government approach, including offensive cyber operations. For example, Tim Watts, Australian MP, argued that the government should 'release the hounds'. In his opinion, the world should take a more proactive approach of imposing costs on ransomware groups by using cyber operations to disrupt their activities and deter them from

attacking high-value targets.[2] This was not the only such statement. In 2021, there was considerable development in this area, including the G7 joint statement on ransomware[3] and the international summit of the Counter Ransomware Initiative.[4] The US National Security Agency (NSA) / Cyber Command[5] and Canadian Communications Security Establishment also confirmed 'imposing costs' on ransomware through offensive cyber operations.[6] The governments of the Netherlands[7] and Australia,[8] and the head of UK GCHQ[9] specifically described offensive cyber operations as a response to prominent ransomware threat actors.

Our paper considers these calls for a whole-of-government approach against ransomware as a starting point. We address two distinct research questions:

1. How do offensive cyber operations influence the ransomware ecosystem?
2. What are the vectors of influence and characteristics of the past operations disrupting the ransomware ecosystem?

We believe answering these questions might further the scientific inquiry and understanding of the implications of this novel and important topic. While this issue is important, it is necessary to bear in mind that it is also controversial. We will steer this paper away from an in-depth discussion of the legal repercussions of offensive cyber operations, but that does not mean these issues are non-existent or even marginal.

Ways to disrupt cybercrime have been treated in previous literature, such as Ryan[10] and Collier et al.[11] Their conclusions and suggestions are discussed together with ours in Section 5 (see below).

2 Tim Watts, 'House debates Wednesday, 2 June 2021: Adjournment Cybersecurity', OpenAustralia, 2 June 2021, https://www.openaustralia.org.au/debates/?id=2021-06-02.144.1.

3 G7, 'CARBIS BAY G7 SUMMIT COMMUNIQUÉ', G7UK, 13 June 2021, https://www.g7uk.org/wp-content/uploads/2021/06/Carbis-Bay-G7-Summit-Communique-PDF-430KB-25-pages-5.pdf.

4 White House, 'Joint Statement of the Ministers and Representatives from the Counter Ransomware Initiative Meeting October 2021', White House, 14 October 2021, https://www.whitehouse.gov/briefing-room/statements-releases/2021/10/14/joint-statement-of-the-ministers-and-representatives-from-the-counter-ransomware-initiative-meeting-october-2021/.

5 Julian E. Barnes. 'U.S. Military Has Acted Against Ransomware Groups, General Acknowledges', *New York Times*, 5 December 2021, https://www.nytimes.com/2021/12/05/us/politics/us-military-ransomware-cyber-command.html.

6 Alex Boutilier, 'Canadian spy agency targeted foreign hackers to "impose a cost" for cybercrime', GlobalNews, 6 December 2021, https://globalnews.ca/news/8429008/canadian-spy-agency-targets-cybercrime/.

7 Catalin Cimpanu, 'Netherlands can use intelligence or armed forces to respond to ransomware attacks', Record, 7 October 2021, https://therecord.media/netherlands-can-use-intelligence-or-armed-forces-to-respond-to-ransomware-attacks/.

8 Australian Government, 'Ransomware action plan', 2021, https://www.homeaffairs.gov.au/cybersecurity-subsite/files/ransomware-action-plan.pdf.

9 GCHQ, 'Director GCHQ speaks at The Cipher Brief Annual Threat Conference', 25 October 2021, https://www.gchq.gov.uk/speech/cipher-brief.

10 Matthew Ryan, *Ransomware Revolution: The Rise of a Prodigious Cyber Threat* (Cham: Springer International Publishing, 2021).

11 Collier et al., 'Cybercrime Is (Often) Boring', 10 June 2020, 9, https://doi.org/10.17863/CAM.53769.

Research on the ransomware ecosystem and cyber operations has the following principal challenges: (i) it is difficult to discern the impact of cyber operations on ransomware-as-a-service (RaaS) based on other variables, including law enforcement disruption, or intergroup, intragroup, and market dynamics, (ii) the ransomware ecosystem is generally very dynamic, with groups disappearing and reappearing as the situation changes,[12] and (iii) the data on changes in the RaaS landscape from public sources has significant observability limitations,[13] and generally did not undergo a rigorous peer-review process. Despite our best effort, these challenges limit our exploratory research and potentially our findings as well.

In Section 2 of this paper, we provide essential characteristics of a contemporary ransomware-as-a-service ecosystem, including factors that make the environment resistant to traditional law enforcement activities. Section 3 contains three case studies of ransomware actors (DarkSide/BlackMatter, REvil, Conti), which we consider highly representative of the current ecosystem and the effect offensive cyber operations had on them. Section 4 presents some initial observations about the influence of offensive cyber operations on ransomware and outlines some best practices available to governments facing large-scale cybercrime groups. In Section 5, we discuss our findings and present additional discussion of the topic. Section 6 concludes the paper.

2. THE RANSOMWARE-AS-A-SERVICE ECOSYSTEM

Over the last few years, the dominant business model of the ransomware ecosystem was RaaS, in which ransomware developers offer malware platforms based on the affiliate model.[14] This contributed to the development of 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, even negotiation and customer support personnel.[15] Despite RaaS being the most prevalent model, some ransomware was privately operated or available for one-time purchase with bundled infection vectors on forums.[16]

12 'ENISA Threat Landscape 2021', Report/Study, ENISA, 41–42, https://www.enisa.europa.eu/publications/enisa-threat-landscape-2021.

13 E.g. Allan Liska, 'Are Ransomware Attacks Slowing Down? It Depends on Where You Look', *Recorded Future*, 20 December 2021. https://www.recordedfuture.com/are-ransomware-attacks-slowing-down/.

14 'ENISA Threat Landscape 2021'.

15 'Ransomware Gangs Are Starting to Look Like Ocean's 11', Kela, 8 July 2021, https://ke-la.com/ransomware-gangs-are-starting-to-look-like-oceans-11/; Catalin Cimpanu, 'This Chart Shows the Connections between Cybercrime Groups', ZDNet, https://www.zdnet.com/article/this-chart-shows-the-connections-between-cybercrime-groups/.

16 Kellyn A. Wagner Ramsdell and Kristin E. Esbeck, *Evolution of Ransomware* (The MITRE Corporation, 2021), 6; 'ENISA Threat Landscape 2021'.

The overall impact of the ransomware ecosystem is difficult to estimate due to the lack of complete data. It is assessed as significant and still rising. US Treasury identified $5.2 billion in Bitcoin transactions (one of the payment methods) potentially tied to the top ten ransomware variants between 2018 and 2021. The amount of suspected ransomware transactions in just the first half of 2021 surpassed the total for 2020.[17] Remediation costs (including business downtime, lost orders, people time and device costs) were estimated to be ten times higher. Even paying victims do not recover all data.[18]

To further the incentives for victims to pay, the majority of RaaS started to use the 'double extortion' tactic of simultaneously exfiltrating and encrypting victim data. Ransomware actors run data leak sites where they publish the exfiltrated data of victims that had not paid. Sometimes the data of paying victims are posted or never deleted. Additional tactics were used to pressure victims and increase the likelihood of payment, such as calls to journalists or business partners and denial of service attacks.[19]

Based on the available information, between 2020 and 2021, RaaS grew significantly compared to previous years. A total of 21 new affiliate programs and 28 data leak sites appeared, publishing the data of 2,371 companies – a 935% increase over the previous year. The total number of victims was likely larger by a factor of ten.[20] The number of initial access brokers and offers tripled. However, only five access brokers were responsible for substantial shares of sales.[21] Brokers and affiliates frequently changed allegiances or simultaneously worked for several ransomware groups. Security researchers also identified overlaps between some strains.[22] These data suggest the RaaS ecosystem might be (i) more centralized and less numerous than it seems from the overall impact, and (ii) this impact is likely unevenly distributed – meaning that the majority of initial access and ransomware attacks seem to be conducted by a minority of the ecosystem. This hypothesis is based on limited available data, which hampers its scientific validity. Nevertheless, the existing data does reduce uncertainty.

Several factors played a major role in RaaS's evolution to the current scale of impact by undermining existing law enforcement agencies (LEA) mechanisms and efforts.

17 Catalin Cimpanu, 'US Treasury Said It Tied $5.2 Billion in BTC Transactions to Ransomware Payments', *Record*, 15 October 2021, https://therecord.media/treasury-said-it-tied-5-2-billion-in-btc-transactions-to-ransomware-payments/.

18 Group-IB, 'Corporansom', *Hi-Tech Crime Trends 2021/2022* (2021), 45, https://www.group-ib.com/resources/threat-research/2021-reports.html.

19 Dmitry Smilyanets, '"I Scrounged through the Trash Heaps... Now I'm a Millionaire:" An interview with REvil's Unknown', *Record*, 16 March 2021, https://therecord.media/i-scrounged-through-the-trash-heaps-now-im-a-millionaire-an-interview-with-revils-unknown/.

20 Group-IB, 'Corporansom', 7–8.

21 Group-IB. 'The Sale of Access to Corporate Networks', *Hi-Tech Crime Trends 2021/2022*, 8, 63–64, https://www.group-ib.com/resources/threat-research/2021-reports.html.

22 E.g. Chainalysis, 'Ransomware 2021: Critical Mid-Year Update', July 2021, https://go.chainalysis.com/rs/503-FAP-074/images/Ransomware-2021-update.pdf.

A. Underground Forums and Marketplaces

Underground forums and marketplaces were websites usually hosted on 'DarkNet' (anonymity-enhancing overlay networks requiring access via specific software, e.g. Tor). They helped lower the entry bar, and provided social and market infrastructure for cybercrime communities, including advertising, sales of initial accesses, recruitment and exchange of information, intrusion tools, and expertise. While these activities also happen elsewhere – for example, in private messaging apps[23] – the advantage of forums and marketplaces are scale, accessibility, inherent trust, and reputation mechanisms, such as limited or invite-only access, escrow services, 'karma' systems based on activity (e.g., number of posts, transactions, cryptocurrencies deposited) or user recommendations.[24]

Forum and marketplace presence also help criminals' reputation and branding, which are important factors both for affiliates and victims. Ransomware with a higher reputation is more likely to attract affiliates, increasing returns on investment into development work and supporting services, such as managing infrastructure or software (malware) development. Public awareness and perception of a given actor's reliability is an important factor in determining whether a victim decides to pay.[25] However, reputation as the core business requirement clashes with the need for anonymity, as it might draw unwanted attention from the media, security researchers, threat intelligence companies, and LEAs.

B. Cryptocurrencies

Cryptocurrencies (also known as virtual or digital currencies) are another enabling factor of modern RaaS. They have been the primary channel capable of moving millions of dollars outside of common government oversight.[26] Despite the public ledger of the most popular cryptocurrencies, criminals used methods that complicate the tracking of payments, such as integrating cryptocurrency mixers into affiliate dashboards to facilitate laundering. The increasing popularity of privacy-enabled cryptocurrencies (e.g. Monero) was an additional factor.[27]

Despite many countries banning or regulating cryptocurrencies,[28] worldwide adoption

23 Caitlin Huey, David Lienbenger, and Dmytro Korzhevyn, 'Translated: Talos' Insights from the Recently Leaked Conti Ransomware Playbook', *Talos Intelligence*, 2 September 2021, http://blog.talosintelligence.com/2021/09/Conti-leak-translation.html.

24 Ryan, 'Ransomware Revolution', 55; Collier et al., 'Cybercrime Is (Often) Boring'.

25 Ryan, 'Ransomware Revolution', 51; Dmitry Smilyanets, 'An Interview with LockBit', *Record*, 26 October 2021, https://therecord.media/an-interview-with-lockbit-the-risk-of-being-hacked-ourselves-is-always-present/.

26 Ryan, 'Ransomware Revolution', 150.

27 Intel471, 'How Cryptomixers Allow Cybercriminals to Clean Their Ransoms', 15 November 2021, https://intel471.com/blog/cryptomixers-ransomware; Chainalysis, 'Ransomware 2021: Critical Mid-Year Update'.

28 Brian Newar, 'The Number of Countries Banning Crypto Has Doubled in Three Years', Cointelegraph, 4 January 2022, https://cointelegraph.com/news/the-number-of-countries-banning-crypto-has-doubled-in-three-years.

is growing.[29] This leaves states with complicated and 'uneven global implementation' of anti-money laundering laws, subject also to geopolitical considerations.[30]

C. The Geopolitical Considerations and Enforcement Gap

The geopolitical considerations and enforcement gap also played a significant role. Ransomware attacks were predominantly aimed at North American and European targets.[31] Multiple sources described most ransomware attacks as originating from cybercriminals in Russia and other Commonwealth of Independent States (CIS) countries[32] – 15 of the 25 most important ransomware groups in mid-2021 were believed to be based there.[33] In 2021, the share of funds extorted by ransomware strains based in the CIS countries grew further. Evidence suggests that these countries were unwilling to intervene as long as threat actors followed basic precautions regarding local targets or helped state intelligence and LEAs.[34]

Although Russian-language RaaS actors currently dominate the market, it is not a case of a single jurisdiction. Even in actively prosecuting countries, some ransomware groups could still operate.[35] Similarly, an overlap between state and criminal actors was identified elsewhere.[36]

29 'The 2021 Geography of Cryptocurrency Report: Analysis of Geographic Trends in Cryptocurrency Adoption and Usage', Chainalysis, October 2021, https://go.chainalysis.com/rs/503-FAP-074/images/Geography-of-Cryptocurrency-2021.pdf.

30 Compare countering illicit finance in the White House. 'Joint Statement of the Ministers and Representatives from the Counter Ransomware Initiative Meeting'.

31 Group-IB, 'Corporansom', 49.

32 Armenia, Azerbaijan, Belarus, Georgia, Kazakhstan, Kyrgyzstan, Moldova, Russia, Tajikistan, Turkmenistan, Ukraine, and Uzbekistan.

33 Some examples: Chainalysis, 'Ransomware 2021'; Andrew E. Kramer, Michael Schwirtz, and Anton Troianovski, 'Secret Chats Show How Cybergang Became a Ransomware Powerhouse', *The New York Times*, 29 May 2021, https://www.nytimes.com/2021/05/29/world/europe/ransomware-russia-darkside.html; Smilyanets, 'I Scrounged through the Trash Heaps... Now I'm a Millionaire'; Smilyanets 'An interview with BlackMatter', Record, 2 August 2021, https://therecord.media/an-interview-with-blackmatter-a-new-ransomware-group-thats-learning-from-the-mistakes-of-darkside-and-revil/.

34 Kramer, Schwirtz, and Troianovski, 'Secret Chats Show How Cybergang Became a Ransomware Powerhouse'; Chainalysis, 'Ransomware 2021; Critical Mid-Year Update'; Joe Tidy. 'Evil Corp: 'My hunt for the world's most wanted hackers'. BBC. 17 November 2021. https://www.bbc.com/news/technology-59297187; Canadian Centre for Cyber Security, 'Ransomware Threat in 2021', Cyber Threat Bulletin, 2021, https://cyber.gc.ca/sites/default/files/2021-12/Cyberransomware-update-threat-bulletin_e.pdf.

35 Catalin Cimpanu, 'The FBI Believes the HelloKitty Ransomware Gang Operates out of Ukraine', *Record*, 15 December 2021, https://therecord.media/the-fbi-believes-the-hellokitty-ransomware-gang-operates-out-of-ukraine/.

36 'Three North Korean Military Hackers Indicted in Wide-Ranging Scheme to Commit Cyberattacks and Financial Crimes Across the Globe', US Department of Justice, 17 February 2021, https://www.justice.gov/opa/pr/three-north-korean-military-hackers-indicted-wide-ranging-scheme-commit-cyberattacks-and; Plan Fraser et al., 'APT41: A Dual Espionage and Cyber Crime Operation', *Mandiant*, 7 August 2019, https://www.mandiant.com/resources/apt41-dual-espionage-and-cyber-crime-operation; 'A Second Iranian State-Sponsored Ransomware Operation "Project Signal" Emerges', *Flashpoint*, 30 April 2021, https://www.flashpoint-intel.com/blog/second-iranian-ransomware-operation-project-signal-emerges/; Smilyanets, 'An Interview with LockBit'; Canadian Centre for Cyber Security, 'Ransomware Threat in 2021'.

Cybercriminals in bulletproof jurisdictions can often be easily tracked by journalists and continue to lead their lives without any disruption. For example, Maksim Yakubets, who is connected to several cybercrime, ransomware, and state intelligence organizations,[37] or Yevgeniy Polyanin, connected to the RaaS REvil.[38]

3. NOTABLE DISRUPTIONS IN THE RANSOMWARE ECOSYSTEM IN 2021

There has been a significant increase in cyber and real-world disruption to ransomware groups, and the second half of 2021 in particular was an active period. Among notable events were the takedown of NetWalker and Maze/Egregor and arrests of more than 30 persons for ransomware crimes and related money laundering, including sanctions on crypto exchanges. Other significant groups shut down voluntarily, such as Babuk or Avaddon.[39]

We have selected three case studies related to the influence of cyber operations on RaaS in 2021. All groups analyzed below were Russian-speaking and likely located in bulletproof jurisdictions.

A. DarkSide/BlackMatter

DarkSide ransomware emerged in August 2020. Pivoting from payment card crime and brief use of REvil ransomware, it started its own RaaS program.[40]

DarkSide stopped operating in May 2021 after an attack on the Colonial Pipeline Company, resulting in temporarily halted pipeline operations, which disrupted the oil supply chain for most of the eastern US.[41] DarkSide later announced that it lost control of the public part of its infrastructure – blog, payment and content delivery network servers 'at the request of law enforcement agencies'. It claimed loss of cryptocurrency and closed the affiliate program 'due to pressure from the US'. Anonymous US officials responded that military or LEAs were not behind the disruption, despite

37 Joe Tidy, 'Evil Corp: "My Hunt for the World's Most Wanted Hackers"', *BBC News*, 17 November 2021, https://www.bbc.com/news/technology-59297187.

38 Will Stewart, 'EXCLUSIVE: REvil "Super-Hacker" Wanted by FBI for "Using Ransomware to Fleece Millions of Dollars" from Americans Is Unmasked by DailyMail.Com in His Plush Hideout in Siberia as Kremlin Turns Blind Eye', *Mail Online*, 28 November 2021, https://www.dailymail.co.uk/news/article-10251531/Russian-hacker-wanted-FBI-using-ransomware-fleece-millions-dollars-unmasked.html.

39 'ENISA Threat Landscape 2021', 40–42.

40 'Ransomware Profile: DarkSide', Emsisoft, Security Blog, 11 May 2021, https://blog.emsisoft.com/en/38577/ransomware-profile-darkside/.

speculations related to the US 780th Military Intelligence Brigade, an offensive cyber capabilities unit.[42]

Later, much of the cryptocurrency ransom from the Colonial Pipeline attack was recovered by US Cyber Command, NSA, and LEA efforts, likely obtaining private keys to Bitcoin wallets. The representative of BlackMatter (rebranded DarkSide) confirmed this in an interview.[43] However, US agencies have not disclosed how they seized the cryptocurrencies, indicating that it took place in a cyber operation.

In July 2021, DarkSide secretly rebranded to BlackMatter and continued operation.[44] BlackMatter has also designated a list of industries not targeted, which reflected critical infrastructure sectors declared as off-limits by US President Biden in a meeting with Russian President Putin.[45] However, BlackMatter was still deployed against victims in these sectors in cases where BlackMatter's definition of 'criticality' differed.[46]

In November 2021, BlackMatter had again announced a shutdown, withdrew cryptocurrency and deactivated accounts, describing 'certain unsolvable circumstances associated with pressure from the authorities'.[47] Although coincidental to the announcement of a multinational cyber operation against REvil, the discovery of a flaw in BlackMatter ransomware allowing decryption of previous victims,[48] ransomware-related arrests of 12 individuals,[49] and a visit by the CIA director to

41 Ibid.

42 Michael Schwirtz and Nicole Perlroth, 'DarkSide, Blamed for Gas Pipeline Attack, Says It Is Shutting Down', *New York Times*, 14 May 2021, https://nytimes.com/2021/05/14/business/DarkSide-pipeline-hack.html; Ellen Nakashima, 'U.S. Government Denies Disrupting Russian Ransomware Ring That Hacked Colonial Pipeline', *Washington Post*, 19 May 2021, https://www.washingtonpost.com/business/2021/05/19/darkside-hack-colonial-cyber-command/.

43 Ellen Nakashima and Dalton Bennett, 'A ransomware gang shut down after Cybercom hijacked its site and it discovered it had been hacked', *Washington Post*, 3 November 2021, https://www.washingtonpost.com/national-security/cyber-command-revil-ransomware/2021/11/03/528e03e6-3517-11ec-9bc4-86107e7b0ab1_story.html; 'Department of Justice Seizes $2.3 Million in Cryptocurrency Paid to the Ransomware Extortionists Darkside', US Department of Justice, 7 June 2021, https://www.justice.gov/opa/pr/department-justice-seizes-23-million-cryptocurrency-paid-ransomware-extortionists-darkside; Smilyanets, 'An Interview with BlackMatter'.

44 Wosar, 'Hitting the BlackMatter Gang Where It Hurts: In the Wallet', Emsisoft Security Blog, 24 October 2021, https://blog.emsisoft.com/en/39181/on-the-matter-of-blackmatter/; Smilyanets, 'An Interview with BlackMatter'; Canadian Centre for Cybersecurity, 'The Ransomware Threat in 2021'.

45 Wosar, 'Hitting the BlackMatter Gang Where It Hurts'.

46 Canadian Centre for Cybersecurity, 'The Ransomware Threat in 2021'; Elizabeth Montalbano, 'BlackMatter Strikes Iowa Farmers Cooperative, Demands $5.9M Ransom', https://threatpost.com/blackmatter-strikes-iowa-farmers-cooperative-demands-5-9m-ransom/174846/.

47 Jonathan Greig, 'BlackMatter Ransomware to Shut down, Affiliates Transferring Victims to LockBit', ZDNet, 3 November 2021, https://www.zdnet.com/article/blackmatter-ransomware-to-shut-down-affiliates-transferring-victims-to-lockbit/.

48 Wosar, 'Hitting the BlackMatter Gang Where It Hurts'.

49 '12 Targeted for Involvement in Ransomware Attacks against Critical Infrastructure', Europol, 2021, https://www.europol.europa.eu/media-press/newsroom/news/12-targeted-for-involvement-in-ransomware-attacks-against-critical-infrastructure.

Russia to discuss, among other things, ransomware,[50] no connection to these events was confirmed.

B. REvil

REvil[51] was first seen in 2019 and soon became one of the top threat actors, working with reportedly 60 affiliates at the peak.[52] Reacting to the disruption of another DarkSide in May 2021, REvil announced the end of advertisement and said it would 'go private', cooperating with only a small group of known and trusted affiliates.[53] Later in 2021, REvil itself would be disrupted in one of the publicly known examples of joint LEAs, military and intelligence agencies' cooperation against ransomware.

After committing its most notable attack in July 2021, the group's leader disappeared from cybercrime forums, and sites maintained by the group became inaccessible. The outage was caused by a cyber operation of an undisclosed government cooperating with the US, permitting the FBI to access servers and private keys subsequently used to create decryption tools. The operation made the ringleaders afraid of identification and arrest and led to the pre-emptive shutdown of infrastructure prior to the planned FBI cyber disruption.[54] It is yet unclear whether Russia was the cooperating government acting based on US-provided information.[55] However, Serhii Demediuk, the Deputy Secretary of Ukraine's National Security and Defence Council, stated in a January 2022 public interview that the July operation was conducted by Russian special services.[56] The significance of this step is discussed below, together with events from January 2022.

In September 2021, REvil briefly resurfaced, only to shut down again in October when the gang's representative announced a breach: a third party had apparently compromised REvil infrastructure and hijacked traffic to disrupt the platform and identify REvil members. The group shut down its infrastructure and discontinued

50 Anton Zverev and Maria Tsvetkova, 'CIA director brings up Russian hackers at talks in Moscow – sources', Reuters, 3 November 2021, https://www.reuters.com/business/cop/cia-director-brings-up-russian-hackers-talks-moscow-sources-2021-11-03/.

51 Also known as Sodinokibi and GandCrab.

52 Smilyanets, 'I Scrounged through the Trash Heaps... Now I'm a Millionaire'.

53 Catalin Cimpanu, 'Darkside ransomware gang says it lost control of its servers & money a day after Biden threat', *Record*, 14 May 2021, https://therecord.media/DarkSide-ransomware-gang-says-it-lost-control-of-its-servers-money-a-day-after-biden-threat/.

54 Victoria Kivilevich, 'Will the REvil Story Finally Be Over?', Kela, 25 October 2021, https://ke-la.com/will-the-revils-story-finally-be-over/; 'FBI Held Back Ransomware Decryption Key from Businesses to Run Operation Targeting Hackers', *Washington Post*, 21 September 2021, https://www.washingtonpost.com/national-security/ransomware-fbi-revil-decryption-key/2021/09/21/4a9417d0-f15f-11eb-a452-4da5fe48582d_story.html; Ellen Nakashima and Dalton Bennett, 'A ransomware gang shut down after Cybercom hijacked its site and it discovered it had been hacked'.

55 Ibid; Martin Matishak, '"Too early to tell" if Russia has cracked down on ransomware gangs, Nakasone says', *Record*, 3 November 2021, https://www.therecord.media/too-early-to-tell-if-russia-has-cracked-down-on-ransomware-gangs-nakasone-says/.

56 Dmitry Smilyanets, 'A top Ukrainian security official on defending the nation against cyber attacks', *Record*, 18 January 2022, https://therecord.media/a-top-ukrainian-security-official-on-defending-the-nation-against-cyber-attacks/.

operations to avoid compromise of their identities. A multi-country operation of law enforcement and intelligence agencies, including US Cyber Command, FBI, and Secret Service, was responsible for the breach, securing access in July after REvil restored infected systems from backups.[57]

Later five REvil affiliates were arrested, and the FBI seized some ransomware payments from a cryptocurrency exchange and affiliate wallet.[58] Information on how they were seized was not released, suggesting that the FBI gained access to a wallet's private key or secret passphrase during the operation.[59] One of the five affiliates, Yevgeniy Polyanin, was indicted. Journalists found out that he was living in Barnaul, Russia, apparently with impunity.[60]

In January 2022, the Russian Federal Security Service (FSB) arrested 14 alleged REvil members based on information and request from US authorities from mid-2021. Eight[61] of those arrested were charged with illicit money laundering/control (cryptocurrencies) and included a person (both REvil and BlackMatter affiliate) responsible for the Colonial Pipeline attack. Together with arrests, millions in cash and luxury cars were also seized. However, no names of the group's leaders and their online handles were released, nor is it known whether REvil ringleaders like Yevgeniy Polyanin are among the arrested. Judging from the timing, many experts also assessed this act to be more tied to the Ukraine crisis than a genuine ransomware crackdown.[62] Cybercriminals themselves predominantly perceived it as a publicity stunt aimed at a formal demonstration of political intent by targeting low-tier members (not developers or 'pentesters') of an already defunct group, not as a policy change. The fact that the charges omitted hacking was also seen as a signal of a long-term process of establishing stronger state control over the ransomware market and overall flow of cryptocurrencies.[63]

57 Ellen Nakashima and Dalton Bennett, 'A ransomware gang shut down after Cybercom hijacked its site and it discovered it had been hacked'; David E. Sanger, 'Russia's Most Aggressive Ransomware Group Disappeared. It's Unclear Who Made That Happen', *New York Times*, 13 July 2021, https://www.nytimes.com/2021/07/13/us/politics/russia-hacking-ransomware-revil.html; Joseph Menn and Christopher Bing, 'EXCLUSIVE Governments Turn Tables on Ransomware Gang REvil by Pushing It Offline', *Reuters*, 21 October 2021, https://www.reuters.com/technology/exclusive-governments-turn-tables-ransomware-gang-revil-by-pushing-it-offline-2021-10-21/. See also Kivilevich, 'Will the REvil Story Finally Be Over?'

58 Dmitri Alperovitch and Ian Ward, 'REvil Is Down—For Now', Lawfare, 16 November 2021, https://www.lawfareblog.com/revil-down-now.

59 Lawrence Abrams, 'FBI Seized $2.3M from Affiliate of REvil, Gandcrab Ransomware Gangs', BleepingComputer, 30 November 2021, https://www.bleepingcomputer.com/news/security/fbi-seized-23m-from-affiliate-of-revil-gandcrab-ransomware-gangs/.

60 Stewart, 'EXCLUSIVE'.

61 Jonathan Greig, 'Moscow court charges 8 alleged REvil ransomware hackers', 15 January 2022, https://www.zdnet.com/article/moscow-court-charges-8-revil-ransomware-hackers/.

62 Brian Krebs, 'At Request of U.S., Russia Rounds Up 14 REvil Ransomware Affiliates', Krebs on Security, 14 January 2022, https://krebsonsecurity.com/2022/01/at-request-of-u-s-russia-rounds-up-14-revil-ransomware-affiliates/.

63 Yelisey Boguslavkiy, 'Storm in "Safe Haven": Takeaways from Russian Authorities Takedown of REvil', AdvIntel, 14 January 2022, https://www.advintel.io/post/storm-in-safe-haven-takeaways-from-russian-authorities-takedown-of-revil.

C. Conti

Conti[64] was first observed in early 2020 and became one of the main groups in 2021.[65] Conti relied on building centralized, highly organized, and skilled teams with a division of labour involving tens of full-time members. It also closely cooperated with criminals behind the Trickbot and Emotet botnets.[66] Affiliates were recruited mostly from forums or through cybercriminal contacts. In both cases, affiliates needed references from recognized cybercriminals before being hired. RaaS owners also often provided infrastructure, exploitation tools, and manuals to support affiliate intrusions.[67]

In August 2021, a dissatisfied Conti affiliate leaked detailed hacking manuals, and technical guides, and published IP addresses of C2 servers and details about two affiliates. Like legitimate businesses, RaaS programs could be impacted through insider threats stemming from human resources.[68]

In November 2021, a security firm breached Conti's payment/recovery services server by exploiting a vulnerability in the access portal. For more than a month, the company collected information from the breached system. Among the information obtained were the real IP and location of the server, several victim chat sessions, login credentials for the server, login credentials for storage service accounts used to extort victims' data, and Bitcoin wallet addresses. Important findings also included IP addresses connecting to the server. However, they all belonged to Tor exit nodes and likely could not be used to identify Conti operators.[69] After the company published the report with findings, the server was shut down for more than 24 hours, uncharacteristic for Conti's generally well-managed infrastructure. The company claimed to have shared its findings with Swiss LEAs. However, it is highly unusual to release such a report in a relatively short timeframe after obtaining access, as it might hamper an investigation.[70]

64 Also known as Ryuk or Wizard Spider, active from at least 2018. Definite confirmation is missing.

65 Catalin Cimpanu, 'Conti Ransomware Gang Suffers Security Breach', *Record*, 20 November 2021, https://therecord.media/conti-ransomware-gang-suffers-security-breach/.

66 Yelisey Boguslavskiy and Vitali Kremez, 'Corporate Loader "Emotet"', AdvIntel, 19 November 2021, https://www.advintel.io//post/corporate-loader-emotet-history-of-x-project-return-for-ransomware; Vitali Kremez and Yelisey Boguslavskiy, 'Backup "Removal" Solutions – From Conti Ransomware With Love', AdvIntel, 29 September 2021, https://www.advintel.io//post/backup-removal-solutions-from-conti-ransomware-with-love.

67 Boguslavskiy and Kremez, 'Corporate Loader "Emotet"'; '[Conti] Ransomware Group In-Depth Analysis', Prodaft, 18 November 2021, https://www.prodaft.com/resource/detail/conti-ransomware-group-depth-analysis.

68 Catalin Cimpanu, 'Disgruntled Ransomware Affiliate Leaks the Conti Gang's Technical Manuals', *Record*, 5 August 2021, https://therecord.media/disgruntled-ransomware-affiliate-leaks-the-conti-gangs-technical-manuals/; Group-IB, 'Corporansom', 77–78.

69 Cimpanu, 'Conti Ransomware Gang Suffers Security Breach'; Prodaft, '[Conti] Ransomware Group In-Depth Analysis'.

70 Ibid.

At the end of February 2022, Conti suffered its largest breach yet. Conti released a statement in which it declared readiness to conduct attacks in support of the Russian government and threatened retaliation attacks on anyone opposing it. A Ukrainian security researcher who had previously infiltrated the group then leaked, almost in their entirety, internal communication records and internal software source code. Among the leaked information were the Conti member hierarchy, handles, and emails, details about the group's infrastructure, storage accounts, cryptocurrency wallets, ransomware negotiations, and payment information. Even active remote connection details, login information, and much contextual information were leaked. This also exposed the group's internal defence measures, such as orders to install endpoint detection and response (EDR) tools on every administrator's computer. At the time of writing (March 2022), information was still being investigated by researchers to clarify further details.[71]

As Krebs reported, notable information revealed a detailed relationship between the group leadership and Russian law enforcement, including the fact that it had been tipped off in October 2021 that it was under investigation by US authorities, and members relaying assurance that the investigation would go nowhere on the Russian side and would be closed soon after.[72] This closely followed events surrounding Conti's top competitor REvil (see above). Similarly, logs seemed to indicate a link between Conti and Russian intelligence agencies. According to Christo Grozev, executive director of the investigative journalism group Bellingcat, the chat logs show that members of Conti tried to hack a Bellingcat contributor with a special interest in files on Alexei Navalny and comments related to Russia's internal security service, FSB. Grozev stated that Bellingcat was informed about this attempt earlier with a warning specifically stating that it was being done on behalf of the FSB.[73] No further information confirming this was publicly available at the time of the writing.

Other contents included a relationship with TrickBot and Emotet gangs, including new details on the US NSA/CyberCommand's disruption of TrickBot in September 2020, which comprised of a takeover of Trickbot infrastructure, disconnection of infected systems from command-and-control infrastructure, removal of failsafe recovery mechanisms, and planting of false records into the Trickbot database. Conti leaders were surprised by the extent of the sabotage. For several weeks the group was unable to rebuild its botnet infrastructure and start infecting new systems. Both this event and the law enforcement operation to seize the Emotet botnet prompted the group to reorganise: as a security precaution, personnel were forced to select new credentials,

71 Brian Krebs, 'Conti Ransomware Group Diaries, Part I: Evasion', Krebs on Security, 1 March 2022, https://krebsonsecurity.com/2022/03/conti-ransomware-group-diaries-part-i-evasion/; Brian Krebs, 'Conti Ransomware Group Diaries, Part III: Weaponry', Krebs on Security, 4 March 2022, https://krebsonsecurity.com/2022/03/conti-ransomware-group-diaries-part-iii-weaponry/.

72 Ibid.

73 Corin Faife, 'A ransomware group paid the price for backing Russia', Verge, 28 February 2022, https://www.theverge.com/2022/2/28/22955246/conti-ransomware-russia-ukraine-chat-logs-leaked.

and low-level staff were even dismissed. Despite that, the response to the March 2022 breach took several days, with the group ultimately unable to identify the leaker and forced to destroy its standing infrastructure and attempt to build it anew.[74]

As of this writing (March 2022), the situation was still developing, including the ongoing assessment of leaked information and its implications –and the mid- to long-term effects are still unclear.

4. INITIAL OBSERVATIONS ON THE DISRUPTION OF THE RAAS ECOSYSTEM

Although still a matter of ongoing debate, the above-described events allow us to make initial observations about the viability and impact of 'releasing the hounds' on ransomware programs.

A. Ecosystem-Level Impact

Correlating with increased disruption and LEAs activity, the RaaS ecosystem has changed in the second half of 2021. While the overall frequency of attacks has likely not slowed, there was a change in the most popular strains, targeting, and operating model.

Between July and September 2021 (i.e. after REvil first disruption), LockBit 2.0, Conti, BlackMatter, and Hive RaaS groups dominated the ecosystem with approximately 60% of observed attacks, a stark difference from the previous period. LockBit 2.0 and Hive emerged only shortly before; BlackMatter and REvil later disappeared altogether.[75] The rest comprised many smaller actors without access or resources to obtain the latest ransomware samples. They relied on niche modification of existing hacking tools.[76]

Fewer attacks were observed against critical infrastructure and sensitive sectors (healthcare, education, municipalities), with some actors stating that they would not target these. However, there were exceptions, such as Conti attacking an Australian electricity provider in December 2021.[77] Secondly, groups likely shifted away from

74 Ibid.; Dan Goodin, 'Conti cybergang gloated when leaking victims' data. Now the tables are turned', Ars Technica, 2 March 2022, https://arstechnica.com/information-technology/2022/03/conti-cybergang-gloated-when-leaking-victims-data-now-the-tables-are-turned/.

75 'A Reset on Ransomware: Dominant variants differ from prior years', Intel471, 15 December 2021, https://intel471.com/blog/ransomware-attacks-2021-lockbit-hive-conti-clop-revil-blackmatter.

76 Thibault Seret, cited in 'The Evolution of Ransomware Operations: Latest Trends', Lifars, 22 November 2021, https://lifars.com/2021/11/the-evolution-of-ransomware-operations-latest-trends/.

77 Joseph Menn, 'Ransomware Attack on Australian Utility Claimed by Russian-Speaking Criminals', Reuters, 9 December 2021, https://www.reuters.com/technology/ransomware-attack-australian-utility-claimed-by-russian-speaking-criminals-2021-12-08/.

large targets, which drew attention to mid-sized victims (preferably in the private sector).[78]

Ransomware actors became increasingly disconnected from the wider cybercriminal community because of unwanted attention. The main forums de-platformed RaaS, which prompted a shift to more-private affiliate programs. In 2021, 29 new data leak sites were detected. However, only 12 new affiliate variants appeared on forums, which suggests the remaining were organized through private affiliate programs. Others established new forums or directly used sales of initial access on marketplaces. However, the disruption was broader than forums banning RaaS partnerships, such as software brokers refusing to sell malware to ransomware groups and affiliates left without means and services to disseminate the ransomware payload.[79]

However, it is unclear from public data whether the change was due to cyber disruption activities or increased LEA activity in general, the reaction of ransomware strains to increased attention, increased cybersecurity spending, decreasing cyber insurance payments, or some combination of these factors.

B. Actor-Level Impact

On the level of individual actors, public disruption operations seemed successful. A compromise of a single REvil server, paired with intelligence collection activity, resulted in the gang's decision to take itself 'offline' and not reappear since. DarkSide/BlackMatter has also so far remained disbanded due to allegedly losing access to some of its infrastructure and 'pressure from authorities.'

Other RaaS platforms also reacted to the development; for example, DarkSide operators quickly moved cryptocurrency after REvil. Conti denounced the attack as 'unilateral, extraterritorial and bandit-mugging' and discussed the legality of hacking servers abroad (sic!), in an ironic statement from a cybercriminal group earning millions of dollars by attacking companies abroad.[80] A LockBit 2.0 representative acknowledged that hacking ransomware infrastructure was: 'one of the most effective methods to deal with us; no one is immune from hacking infrastructure with the help of 0-days'.[81] The representative also discussed his belief that 'the NSA has hardware backdoors

78 'Are Ransomware Attacks Slowing Down? It Depends on Where You Look', Record, 20 December 2021, https://www.recordedfuture.com/are-ransomware-attacks-slowing-down/; 'Ransomware Attackers down Shift to "Mid-Game" Hunting in Q3', *Coveware*, 21 October 2021, https://www.coveware.com/blog/2021/10/20/ransomware-attacks-continue-as-pressure-mounts; 'Ransomware Tracker: The Latest Figures [January 2022]', *Record*, 10 November 2022, https://therecord.media/ransomware-tracker-the-latest-figures/.

79 Group-IB, 'Corporansom', 40–44.; Vitali Kremez and Yelisey Boguslavskiy, 'The Rise & Demise of Multi-Million Ransomware Business Empire', AdvIntel, 15 June 2021, https://www.advintel.io/post/the-rise-demise-of-multi-million-ransomware-business-empire.

80 Ivan Righi, 'Ransomware Q3 Roll Up', Digital Shadows, 25 October 2021, https://www.digitalshadows.com/blog-and-research/ransomware-q3-2021-roll-up/.

81 Smilyanets, 'An Interview with LockBit'.

in any server on the planet.'[82] Political reasons and fear of retaliatory measures from the US or Russia were also likely[83] behind the decision of Avaddon ransomware to disband its up-and-coming and very profitable program without an explanation.

C. Cyber Operational Options

As further explained below, ransomware actors face issues similar to their legitimate counterparts: (i) securing their infrastructure is a complex task, (ii) operational security (anonymity) on the internet is difficult over a long period of time, and (iii) brand-building for RaaS is in direct clash with their desire to avoid unwanted attention. All these factors intensify when RaaS activity is scaled and can be leveraged for offensive cyber operations.

a) Securing Infrastructure

On the infrastructure level, RaaS have two options for their public-facing infrastructure (i.e. data leak sites and payment portals): they can either use public services, which is accompanied by a higher risk of takedown, or build a robust Darknet infrastructure required to handle various functions such as communication, payment automation or data transfers. Failure to build sufficient infrastructure for handling exfiltrated data can dissuade buyers and affiliates.[84] RaaS groups also face the same issues as legitimate administrators, including vulnerability management, access control, and human resources issues. As referenced in the cases above, even highly skilled ransomware groups known for well-managed infrastructure can be impacted by vulnerabilities or insider threats.

Inability to secure the infrastructure can make intelligence gathering easier. In addition to the examples described in the cases of REvil and Conti, it is also possible to access affiliate chats, Jabber communication accounts, and cryptocurrency wallet details,[85] or identify specific members. In a similar way, it is possible to leverage insecurities to cause persistent disruption, which imposes monetary and temporary costs on ransomware operators and directly impacts their profitability. Examples of best-case persistent disruption are provided in the examples further below.

b) Operations Security and Anonymity Difficult Over Time

As Brian Krebs has repeatedly shown in the 'Breadcrumbs' investigation series, cybercriminals make operations security mistakes, often due to inadequate separation of virtual identities, credential reuse, or leaked information from hacked cybercriminal

82 Ibid.
83 Kremez and Boguslavskiy, 'The Rise & Demise of Multi-Million Ransomware Business Empire'.
84 For a discussion on data leak sites, see e.g. Righi, 'Ransomware Q3 Roll Up'.
85 '[LOCKBIT] Behind The Lines of LockBit R.a.a.S.', PRODAFT, 18 June 2021, https://prodaft.com/resource/detail/lockbit-behind-lines-lockbit-raas.

forums that reveal secrets or connections. Increased likelihood of mistakes over time enables forum personas to be connected with real-life identities.[86]

While the use of some of this information by LEAs and the justice system can be limited, they are often picked up by journalists.[87] Military and intelligence agencies can use them effectively for intelligence gathering and cyber, influence, and psychological operations.

In some cases, even criminals can be nudged in this direction, for example, through reward offers for information on the identity and location of conspirators and participants in RaaS groups. The US State Department offered significant rewards for information on persons connected to DarkSide and REvil. Experts described this as an effort to impact the environment psychologically and create distrust between criminals.[88]

c) Forum and Marketplaces Level

While not traditionally associated with just ransomware, disruption of the current ecosystem of underground forums and marketplaces would severely impact RaaS programs.[89] Disruption could especially affect initial access sales, recruitment, and advertising, and also hinder the exchange of skills, tools, and information. All are necessary for the current model of RaaS and the development of new tools, exploits, and ransomware variants. Furthermore, disruption affects the overall trust in criminal communities and forces criminals to focus on other means of cooperation and communication. It increases friction and operational expenses. It also lowers profitability.

D. Previous Cyber Disruption of Non-State Actors

For a better overview of aspects of a successful disruption operation, we have also turned to previous success stories of cyber operations aimed at disrupting non-state actors in cyberspace – the 2016 anti-ISIS Operation Glowing Symphony and 2021 Operation Ladybird, disruption of Emotet botnet. While there were numerous botnet disruptions, Operation Ladybird was selected due to its recentness and highly successful outcome. These case studies remain merely analogous and do not specifically describe the disruption of RaaS. However, they showcase factors for consideration such as intensity, scope, range of actors, and techniques involved.

86 Brian Krebs, 'Category Archives: Breadcrumbs', Krebs on Security, 2020–2021, https://www.krebsonsecurity.com/category/breadcrumbs/.

87 See Chapter 2 above.

88 Alperovitch and Ward, 'REvil Is Down—For Now'.

89 Cf. e.g. Anastasia Sentsova, Andrew Mincin, and Yelisey Boguslavskiy, 'Four Scenarios of Attacks on DarkWeb Forums', AdvIntel, 30 March 2021, https://www.advintel.io/post/four-scenarios-of-attacks-on-darkweb-forums-adversarial-perspective-post-incident-analysis.

a) Operation Glowing Symphony

Operation Glowing Symphony was a 2016 joint cyber operation of the US Cyber Command and NSA against the 'Islamic State' terrorist group.[90] It focused on the group's digital recruitment, financial activities, social media, and propaganda infrastructure. It was the first, largest and (at the time) most complex offensive cyber operation publicly acknowledged by the US government.[91]

The operation focused on obtaining access to target infrastructure, intelligence collection, data removal, and disrupting access for group operators and administrators. Continued disruption over at least the next seven months was aimed at psychologically impacting IS personnel, for example, through inducing anger, confusion, and distrust by deliberately causing technical issues and problems that looked like common IT issues (e.g. slow or cut connections, denying access, disrupting content or communication). The secondary effect of this disruption forced some IS operatives to use less secure tools and expose their position, which enabled physical targeting.[92]

While aspects of the operation have been a matter of ongoing debate, IS digital operations were reduced and disrupted at the time of crucial on-the-ground developments. Up to 60% of the IS digital infrastructure never recovered. Media production had significantly lowered quality and delays.[93]

b) Operation Ladybird

In January 2021, a multinational force of LEAs and security researchers, coordinated by Europol and Eurojust, disrupted the Emotet botnet. It was one of the most significant and long-lasting botnets of the past decade, with close ties to ransomware actors (Ryuk/Conti).[94]

Emotet C2 global infrastructure of hundreds of servers controlling 1.6 million bots with redundancy and anti-takedown measures.[95] In a novel approach to botnet disruption, investigators took control of the infrastructure, disrupted it from the inside, redirected infected machines towards controlled infrastructure, and set a clean-up command to

90 Self-proclaimed, not a recognized state.

91 Alan Haji, 'Operation Glowing Symphony (2016)', Cyber Law Toolkit, last modified 4 June 2021, https://cyberlaw.ccdcoe.org/wiki/Operation_Glowing_Symphony_(2016); Dina Temple-Raston, 'How The U.S. Hacked ISIS', NPR, 26 September 2019, https://www.npr.org/2019/09/26/763545811/how-the-u-s-hacked-isis?t=1641204902697; Michael Martelle, 'USCYBERCOM After Action Assessments of Operation GLOWING SYMPHONY', National Security Archive, 21 January 2020, https://nsarchive.gwu.edu/briefing-book/cybervault/2020-01-21/uscybercom-after-action-assessments-operation-glowing-symphony.

92 Ibid.

93 Ibid.

94 Boguslavskiy and Kremez, 'Corporate Loader "Emotet"'.

95 'World's most dangerous malware EMOTET disrupted through global action', Europol, 18 November 2021, https://www.europol.europa.eu/media-press/newsroom/news/world%E2%80%99s-most-dangerous-malware-emotet-disrupted-through-global-action; James Shank, 'Taking Down Emotet', Team Cymru, 27 January 2021, https://team-cymru.com/blog/2021/01/27/taking-down-emotet/.

all hosts regardless of jurisdiction. The operation included lawful physical access and arrests.[96]

While activities necessary to disrupt a botnet differ from disrupting RaaS, this case shows interesting aspects of influencing infrastructure beyond the reach of law enforcement collaboration. In some countries, Emotet was not illegal unless it targeted citizens. Some hosting providers also might have had ties to the criminal enterprise and alerting them about upcoming activity could have warned malicious actors. As a result, network operators in some jurisdictions were enrolled through informal, peer-to-peer contacts. Similarly, parties judged unlikely to cooperate (or untrustworthy) were sidestepped, as operators blocked traffic to the subset of Emotet servers based in these countries. That allowed the operation to take over most of the C2 and servers unable not under control.[97] Dutch police also used a cyber operation to penetrate Emotet's infrastructure, including discovering and disrupting infrastructure backups.[98]

Ten months later, helped by Conti/Ryuk, Emotet returned with updated, more secure code and infrastructure.[99] Despite this, the operation was arguably successful as it stopped a major initial access vector for other malware for a significant time. It also serves as a model of multinational cooperation between LEAs, private companies, and engaged individuals.

c) Best Practice

Formulation of best practices based on two operations is not optimal. However, these operations present us with possible trends in offensive cyber operations. They have shown us the importance of careful planning based on a thorough understanding of the environment and importance of continuous pressure over the one-off or short-term disruption.

Both operations heavily relied on cooperation. Multinational and multilateral (public and private sector) cooperation is a key element in ensuring that geographically dispersed malicious cyber activities are efficiently disrupted or even suppressed. This cooperation can occur through existing mechanisms or through ad hoc channels.

Understanding the environment also presents a challenge but seems to pay off. The environment consists not only of criminal actors, but also of LEAs or private companies. Both sides of the landscape need to be understood properly, not only to efficiently disrupt the ongoing illicit activity but also to assess whom to bring on

96 Ibid.
97 Ibid.
98 Andy Greenberg, 'Cops Disrupt Emotet, the Internet's "Most Dangerous Malware"', *Wired*, 27 January 2021, https://wired.com/story/emotet-botnet-takedown/; Sean Lyngaas, 'FBI leaned on Dutch cops' hacking in Emotet disruption', *Cyberscoop*, 5 February 2021, https://www.cyberscoop.com/fbi-emotet-dutch-takedown-cybercrime/.
99 Boguslavskiy and Kremez, 'Corporate Loader "Emotet"'.

board in terms of cooperation to ensure the success of the operation. Additionally, both operations have shown that complete success (i.e. disbanding the target actor) is highly unlikely. Setting a realistic goal is important and defining levels of disruption in terms of heightened operational costs or lowered efficiency of criminal actors should be considered. While moderate goals might be perceived as insufficient, achieving such a goal might already have a significant impact.

5. DISCUSSION

Collier et al.[100] describe how 'crackdowns' can unite communities and give them a common sense of struggle and persecution. They argue that making the repetitive, tedious, and 'boring' work of cybercrime administrative workers even more boring through disruption presents a strategy more viable than, for example, legitimate takedowns through network providers. The authors then continue to discuss messaging around cybercrime as a pathway to influencing the behaviour of low-level workers. According to available data from both the threat intelligence and blockchain analysis companies, the RaaS ecosystem seems unevenly distributed. Some more skilled and better-organized actors are responsible for most initial access sales, ransomware deployments, and profits. We believe this is in line with our findings that causing disruptions and leveraging security difficulties that even legitimate businesses are facing can significantly influence the environment. Imposing additional costs on RaaS agents is a way to avoid the unifying effect of 'crackdowns'.

Based on an independent inquiry into new case studies, we have arrived at a similar conclusion as Ryan.[101] However, where Ryan asks whether states need to develop alternative responses and consequences for states that launch or support ransomware attacks,[102] we explored offensive cyber operations directed against these groups as a possible alternative tool. It might not be necessary to develop strategies against nations if some results can be achieved on the level of cybercrime actors.

In this regard, we believe we offer additional insight into this issue, despite the limitations of our research described in Section 1.

A significant number of avenues are open to further research into offensive cyber operations against the RaaS ecosystem. Take for instance issues pertaining to the operational aspects. As mentioned, understanding the environment is crucial. However, disrupting the environment complicates the process of intelligence gathering and observation of the environment. Striking the right balance between the ability to observe and understand and the ability to disrupt is beyond the scope of this paper.

100 Collier et al., 'Cybercrime Is (Often) Boring'.
101 Ryan, 'Ransomware Revolution', 153–156.
102 Ibid., 155.

Secondly, there are the issues inherent to any offensive cyber operation pertaining to collateral damages and inadvertent escalation. Offensive cyber operations against RaaS ecosystem can be misinterpreted by non-participating LEAs or other bodies. Additionally, there is always a risk of an accidental or spillover effect on an unintended target.[103] Again, striking a balance between proactive measures and caution is beyond the scope of this paper and requires further discussion. Third, the mere fact that offensive cyber operations might have a significant influence on the RaaS environment does not necessarily mean these should be widely used. There are other values and approaches to consider (among them relevant legal issues). These can have both international and national consequences. Comprehensive policy options need to be developed accounting for the institutional background of individual states or agencies, including the cost-benefit analysis of conducting offensive cyber operations.

These avenues also delineate logical counterarguments to the viability of our conclusions. The risk of over-militarization of the response to cybercrime is present, and 'releasing the hounds' could open the floodgates if left unchecked. The legality of offensive cyber operations remained completely outside the scope of this paper, but it is imperative that some limits be imposed. Similarly, offensive cyber operations could push ransomware actors away from targets relevant to national security (e.g. critical infrastructure and essential services) towards smaller private companies. The rising number of attacks against lower-level private targets can have a similar cumulative economic impact as large-scale attacks.

Offensive cyber operations are unlikely to be the silver bullet to degrading and eliminating ransomware attacks. However, they can be an important tool in the whole-of-government and multinational efforts. In spite of it all, RaaS is one of the most important cyberthreats for 'every single day' for at least the next several years.[104]

6. CONCLUSION

Due to specific factors, such as (i) self-organization of business relations, skill, and knowledge transfers in the environment of underground forums and marketplaces, (ii) cryptocurrencies, and (iii) geopolitical considerations, the traditional law enforcement approach was not effective in sufficiently disrupting the RaaS ecosystem which in measurable terms grew severalfold between 2019 and 2021. Many industry experts and governmental organizations have lately announced or supported the use of offensive cyber capabilities to disrupt ransomware groups. We subjected this to

103 See e.g. Martin Matishak, 'Pentagon official: 'Open question' if Putin's government can stop hackers', *Record*, 20 October 2021, https://therecord.media/pentagon-officialopen-question-if-putins-government-can-stop-hackers/.

104 Martin Matishak, 'NSA chief predicts U.S. will face ransomware "every single day" for years to come', *Record*, 5 October 2021, https://therecord.media/nsa-chief-predicts-u-s-will-face-ransomware-every-single-day-for-years-to-come/.

an initial inquiry through a study of publicly available information structured into RaaS case studies focusing on the most important ransomware groups (DarkSide/ BlackMatter, REvil, Conti).

Our first research question addressed the influence of offensive cyber operations on the ransomware ecosystem. As described in the case studies, even top RaaS groups can be disrupted through offensive cyber operations. Exploiting vulnerabilities to disrupt systems, social engineering, psychological operations to further internal fault lines, and leaking critical data influences groups' ability to execute their core business function. RaaS groups face similar challenges to cybersecurity and operational security, as do legitimate businesses. Any disruption that imposes additional costs and instils reputation damage on RaaS groups is important. Payment and affiliate portals are critical for current RaaS operations, as downtime or even a lack of positive user experience for both victims and affiliates directly impacts the business model. Case studies have shown that this impact is significant, and offensive cyber operations are a viable part of the whole-of-government approach to tackle this issue. The results of these operations in the past have included the collection of critical intelligence and have had a direct impact on the profitability of operations: imposing infrastructure recovery, internal security costs, loss of reputation, and even increased stress on members, staff dismissals, and groups disbanding altogether.

Our second research question addressed the vectors of influence and characteristics of the past operations disrupting the ransomware ecosystem. Potential disruptions of underground forums and marketplaces impact ransomware advertising and reputation, which hampers recruitment and the willingness of victims to pay. It constrains the ability of RaaS operators to exchange skills, tools, and information, which slows the development and deployment of new exploits and ransomware strains. Furthermore, disruption and compromise affect the overall trust in criminal communities and force RaaS operators to hastily establish new forums or communication methods. This can raise operational expenses, lower profitability, and increase the potential for lower security standards and misconfigurations of alternatives for cybercrime actors, while providing governments with further options to direct criminals to monitored platforms.

As seen `from the examples of Operation Glowing Symphony and Operation Ladybird, successful cyber operations can have a lasting impact on criminal activities. After Operation Glowing Symphony, up to 60% of the Islamic State's infrastructure never recovered. Operation Ladybird left the Emotet botnet severely disrupted for almost a year. Focus on cooperation, knowledge of the environment, and pursuit of reasonable goals is advisable to ensure mid- to long-term disruption and success of offensive cyber operations.

Understandably, many other questions are yet to be addressed in further research (see Section 5 for discussion).

ACKNOWLEDGEMENTS

Jakub Harašta's contribution to this paper was supported by European Regional Development Fund (ERDF) project 'CyberSecurity, CyberCrime and Critical Information Infrastructures Centre of Excellence' (No CZ.02.1.01/0.0/0.0/16_019/0 000822).

Our special thanks go to our families and relatives for their support, motivation, and toleration of our work schedules.

Thank you also to all security researchers and journalists who continue to cover the ransomware topic despite all of us collectively being tired of its volume. Your work is much appreciated. Special mention goes to Patrick Grey and Adam Boileau for their continuous coverage and inspiration.

We are grateful to all network defenders who tirelessly face ransomware and other cyberattacks.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

Third-Party Countries in Cyber Conflict: Understanding the Dynamics of Public Opinion Following Offensive Cyber Operations

Miguel Alberto Gomez
Senior Researcher
Center for Security Studies
Swiss Federal Institute of Technology
Zurich, ZH, Switzerland
miguel.gomez@sipo.gess.ethz.ch

Gregory Winger
Assistant Professor
School of Public & International Affairs
University of Cincinnati
Cincinnati, OH, United States
Gregory.Winger@uc.edu

Abstract: The transnational nature of cyberspace alters the role of third-party countries (TPCs) in international conflict. In the conventional environment, military operations are primarily confined to the boundaries of the combatants or a designated war zone. However, during cyber conflict, operations may occur on the digital infrastructure of states not otherwise involved in the dispute. Within the cybersecurity literature, little is said regarding the role of TPCs who, by virtue of interconnectivity, may find themselves involved in a conflict not of their own making. Consequently, we examine the political and diplomatic hazards inherent in cyber operations that involve these actors. Using a survey experiment fielded in the United Kingdom, we assess the impact of revelations of offensive cyber operations on a TPC population. We test whether prior authorization, existing alliances, and the nature of the target influence public opinion in TPCs following cyber operations conducted within their digital space. We find that while these individuals view these incidents negatively, prior authorization and the involvement of an ally mitigate negativity. Negativity, however, is less affected by target identity.

Keywords: *alliances, public opinion, survey experiment, third parties, cyber operations*

1. INTRODUCTION

In 2016, the United States Cyber Command launched a unique military campaign. Operation Glowing Symphony was a cyber offensive intended to deny, degrade, and disrupt the digital infrastructure of the Islamic State of Iraq and Syria (ISIS) (Temple-Raston 2019). However, before Cyber Command could delete its first ISIS video, the Obama administration was wracked by a bitter internal debate. Although ISIS was the intended target, the group's digital footprint spanned international borders. This meant that while conventional military operations principally occurred inside Iraq and Syria, Glowing Symphony's target list included 35 countries (Nance & Sampson 2017). The question that confounded the Obama administration was not whether to use cyber operations against ISIS but how the US could conduct offensive cyber operations against systems within foreign governments' sovereign territory, including US allies.

The diplomatic dilemma raised by Operation Glowing Symphony highlights the precarious position of third-party countries in the age of cyber conflict. Whereas conventional military operations are physically bounded, cyberspace's transnational and interconnected nature means that cyber operations are geographically dispersed. Consequently, governments and populations who may not otherwise be parties to a dispute may nevertheless play unwitting hosts to digital dogfights waged within their cyber infrastructure. We define such states as third-party countries (TPC) because they are not immediate parties to a conflict but their digital infrastructure is still subject to operations waged between foreign adversaries.

This article explores the political and diplomatic hazards inherent in conducting cyber operations involving TPCs and their implications for international politics. Cyber conflict continues to unfold in the public eye, and elite-centered scholarship is no longer sufficient to understand the emerging political realities of the digital domain. Expanding on recent public opinion scholarship in cybersecurity (Kreps and Schneider 2019; Shandler et al. 2021a; Gomez and Whyte 2021), we employ a survey experiment to discern the consequences of cyber operations involving TPCs. Utilizing the unique approach of experimental subjects serving as bystanders rather than the intended targets, we test whether essential conditions like host government approval, alliance status, and target identity influence public perceptions of the cyber operation itself and overall sentiment towards security partnerships in cyberspace.

We find that positive sentiment toward cyber operations within TPCs is contingent on prior approval by national governments and existing alliances. The obtaining of authorization and the existence of an alliance beforehand results in TPC publics finding benefit in the said action and, overall, being willing to support similar acts in

the future. Consequently, publics are less inclined to punish initiators of the operation than in instances where neither authorization nor an alliance exists. Moreover, we also find that the identity of the intended target (i.e., state versus non-state) does not alter public perceptions, which remain contingent on operational considerations (i.e., authorization) and strategic relationships (i.e., the existence of an alliance) with the initiating state.

These findings reveal the influence of cyber operations disclosure on public opinion within TPCs. Furthermore, while public preferences are not the sole determinants of foreign policy, they exert a significant influence over elite decision-making, one worthy of further study (Putnam 1988). Furthermore, the article reflects the growing curiosity towards public opinion among cybersecurity scholars interested in policy development vis-à-vis cyberspace.

2. THIRD-PARTY COUNTRIES IN TRANSNATIONAL CONFLICTS

The situation of TPCs has become more fraught with the rise of cyber conflict, but neither their contested status nor potential entanglement in foreign conflicts originated with the digital domain. International conflicts have never been wholly contained to the borders of combatants, and spill over into TPCs is a regular feature of militarized international disputes (see Siverson and Starr 1991; Buhaug and Gleditsch 2008; Gleditsch, Salehyan, and Schultz 2008). Indeed, overfly privileges, transit agreements, and basing issues are emblematic of the role TPCs can play in international disputes without being party to the conflict themselves (e.g., Mason 2010; Pettyjohn and Kavanagh 2016). However, in the 20th century, the issue of TPCs as loci for foreign operations became more pointed thanks to globalization and the growth of transnational forms of conflict. In the 1960s and 1970s, groups like the Popular Front for the Liberation of Palestine and individuals like "Carlos the Jackal" pioneered new tactics like airplane hijackings that exploited a globalized environment and lax security conditions in TPCs, allowing them to attack their primary adversaries (Enders and Sandler 2002; Rapoport 2004; Hughes 2014). Unsurprisingly, governments like Israel began expanding their operations into TPCs (Reeve 2000; Klein 2007). Following the 9/11 terrorist attacks, the Bush administration asserted a unilateral right to conduct counterterrorism operations against targets anywhere in the world (Cronin 2002; Patman 2006). As such, TPCs became integral to the Global War on Terror, with friends and adversarial states alike incorporated into US intelligence and counterterrorism operations (Clarke 2004; Schmitt and Shanker 2011; Schaller 2015; Owen and Maurer 2014). This interventionist approach posed a unique challenge for American partners. While national governments might support

combating international terrorism, how that conflict manifested within their borders or impinged on their sovereignty remained an open question (Acharya 2007; Biswas 2009; Jackson 2007).

As transnational conflict and intelligence contests (Rovner 2020) merge in the digital environment, this increasingly important arena makes TPCs a critical facet within cyber conflict. This assent stems from the uniquely transnational nature of cyberspace, straddling both the physical and virtual worlds (Libicki 2009). Cyberspace may be a domain where information flows unimpeded by national boundaries, but it relies on infrastructure in the physical world and under the jurisdiction of sovereign states. Nevertheless, the ability to remotely create, access, or interact within cyberspace creates an environment whereby actions are not limited by geographic proximity or national jurisdiction. This transnational environment emphasizes the salience of TPCs, since data can be created, stored, or even attacked by adversaries within the TPC networks without either combatant being physically present.

Operation Glowing Symphony encapsulates this dynamic. Although ISIS was primarily based in Iraq and Syria, its digital operations spanned the globe and included servers physically located in the United States, Canada, Belgium, and the Netherlands (Pop and Rasmussen 2018; Nance and Sampson 2017). The US Cyber Command created a list of targets that spanned 35 countries, including US allies, to disrupt this network. Declassified documents indicate that some foreign governments, notably Israel and the Netherlands, were involved in planning the operation in some capacity, but many others were not (Department of Defense 2016). Specifically, the idea of asking the German government for permission to conduct cyber operations on their networks sparked a debate within the Obama administration (Nakashima 2017). Eventually, the administration decided to notify the German government about the operation but to expressly not ask for permission.

Glowing Symphony remains the best-known instance of a TPC ensnared in an ongoing cyber conflict and heralds a growing feature of international competition. In March 2021, General Paul Nakasone, who leads US Cyber Command and the National Security Agency, testified that the US conducted more than two dozen cyber operations to prevent interference in the 2020 US election. These included 11 “hunt forward” operations in nine different countries where US cyber forces worked with partners to address malicious actors (Gazis 2021; Vavra 2020). Amid such trends and the ongoing proliferation of cyber capabilities (DeSombre et al. 2021; Craig 2020), it is essential to understand how these are perceived within TPCs and the political fallout these may produce.

3. ACCOUNTING FOR THIRD-PARTY COUNTRIES

Thus far, the issue of third-party countries in cyber conflict has been conceived as a legal question focusing on the rights and protections afforded to TPCs under the laws of war (Schmitt 2017; Lin 2012). However, these concerns do not exist in a vacuum and cannot be divorced from broader domestic and international political dynamics. Specifically, disclosing covert actions can inflame public sentiment (Otto and Spaniel 2021) and stymie even the soundest geopolitical partnerships (Easley 2014).

This highlights an underlying deficiency in how the phenomenon of cyber conflict is studied. While the literature is expanding rapidly, it is focused on political elites while eschewing the increasingly public nature of cybersecurity. Cyber operations have the potential for visible and possibly substantial effects. The need to understand how publics perceive such threats has become more urgent following episodes like the ransomware attack on the Colonial Pipeline. While the cybersecurity literature mainly approaches the issue from the perspective of elite decision-making, this viewpoint fails to recognize how publics respond to such incidents and the possible influence exerted on policymakers. With cyberspace fast becoming a crucial feature of modern society, threats to and from it shape the public discourse that determines the level of support for specific national and foreign policies.

Recent inquiries into the nexus between cybersecurity and public opinion provide a critical window to examine the political implications of cyber operations within TPCs. Scholarly investigations into cybersecurity's resonance among the public often focus on attribution (Gomez 2019b, 2019a), threat perception (Kreps and Schneider 2019; Kostyuk and Wayne 2021), and support for retaliation (Shandler et al. 2021a; Gross, Canetti, and Vashdi 2017). While these phenomena are relevant, they reflect the immediate reaction of publics in states directly targeted by cyber operations. There is little said concerning the response of third-party populations who are bystanders to cyber operations conducted on or through their digital infrastructure.

A. Public Backlash

Clandestine operations into the territory of sovereign states are inherently controversial, and the public disclosure of these, regardless of their provenance, can trigger a significant outcry (Otto and Spaniel 2021; Smith 2019; Lin-Greenberg and Milonopoulos 2021). We hypothesize that this likely extends into cyberspace with the revelation of foreign cyber operations provoking adverse reactions from TPC publics. However, a distinction must be made between discovering covert cyber operations and assuming their existence as the consequences of either are distinct (i.e., confirmation of such an assumption may lead to a stronger public response).

Interstate conflict in cyberspace is an opaque phenomenon, given that publics have limited access to information on it. However, elites may have greater knowledge of cyber operations, due to their access to resources that often remain undisclosed to keep specific capabilities secret. Consequently, it can be argued that cyber operations come to the attention of publics through media reporting (Jarvis, Macdonald, and Whiting 2017), which, as a function of limited access to information or governmental prerogatives, may not accurately portray the given incident. This has several implications as to how publics may respond.

Setting aside perceived violations of sovereignty, adverse public reactions may be amplified by an underlying sense of dread towards cyberspace (Gomez and Villar 2018; Dunn Cavelty 2013). Given the pervasive lack of expertise among elites and the public alike (Hansen and Nissenbaum 2009) and the rarity of severe cybersecurity incidents, beliefs regarding the consequences of cyber incidents are likely influenced by readily accessible portrayals (i.e., popular media) that do not accurately represent the nature and consequences of these events (Reinhardt 2017). On a related note, elite narratives following high-profile events often rely on analogies (e.g., Cyber Pearl Harbor) ill-suited to reality (Lawson and Middleton 2019). Consequently, publics may assess severity using heuristic devices that fail to capture objective reality (Viscusi and Zeckhauser 2017; Gigerenzer 2006).

H_1. *The disclosure of cyber operations involving the cyber infrastructure and assets of a third-party country increases negative reactions from the general population of that country.*

B. Seeking Permission

We hold that, despite the potential for backlash, adverse reactions to these operations can be assuaged through prior consent from the TPC government. As TPCs can authorize the use of their airspace or waterways for military operations, so too will the TPC's granting of permission for military operations in cyberspace signal both foreknowledge and approval of the operation to its population. Moreover, absent any additional information, authorization from a TPC government could provide an essential cue for elite consensus on the operations, which might ameliorate widespread angst (Kreps 2010).

H_2. *Prior authorization from the government of the third-party country reduces negative reactions from the general population.*

C. Existing Alliances

Establishing relationships between the TPC and the perpetrating state may influence public perceptions when the operation lacks prior authorization. Specifically, absent

more detailed information, the simple status of a perpetrating state as either a "partner" or an "ally" may signal that the operation was not malicious and may even have been in the interests of the TPC. This argument coincides with the value-signaling seen in other alliance arrangements (Tomz and Weeks 2021). Furthermore, it reflects the ability of the public to judge the merits of a foreign policy issue even without elite cues (Kertzer and Zeitzoff 2017).

H_3. *The status of the attacker as an "ally" of the third-party country reduces negative reactions from the general population.*

D. Target Identity

Just as the status of the operation's perpetrator may influence public perceptions, so too may the target's identity. Experimental research suggests that publics are generally amenable to aggressive operations in the context of an ongoing cyber conflict. Shandler et al. (2021a) demonstrate pronounced support for retaliation among publics following a severe cyber operation. However, subsequent research notes that this is contingent on knowledge of the possible costs associated with policy choices, such as the loss of life or the possibility of escalation (Shandler et al. 2021b; Kreps and Schneider 2019).

Consequently, how TPC publics react rests on understanding how state and non-state actors differ regarding cyber capabilities. Considering the lack of cyber expertise among publics (Kostyuk and Wayne 2021), knowledge of such differences rests on cues from both elites and the media. In their analysis of media representation of cybersecurity incidents, Jarvis, Macdonald, and Whiting (2017) note that both the identity and purported skill sets directly influenced the level of concern expressed in articles. While this does not explicitly distinguish between state and non-state actors, elite narratives emphasize the pervasiveness of state-sponsored cyber operations that may result in the formation of beliefs among the public as to how states may respond to such threats (Kreps and Schneider 2019; Shandler et al. 2021a).

H_4. *Negative reactions from the public increase if the intended target of the cyber operation is a state actor.*

4. METHODOLOGY

The article employs an Internet-based between-subject survey experiment with participants from the United Kingdom recruited through Prolific Academic, a crowdsourcing platform. The experiment posits a fictitious scenario involving a cyber operation conducted by an unnamed state actor targeting cyber infrastructure owned

and operated by organizations within the United Kingdom that an unnamed adversary uses to conduct its own cyber operations.

The scenario varies across three fundamental details that function as the experimental treatments. These are *authorization, alliance,* and *actor*. After reading the scenario, participants are asked to consider possible responses to and consequences of the incident. These serve as the outcome (i.e., dependent) variables and are identified as censure, continue, and consequences. Furthermore, the experiment considers participants' underlying beliefs, cybersecurity knowledge, and demographic attributes, as these may provide alternative explanations for their preferences.

A. Treatment and Outcome Variables

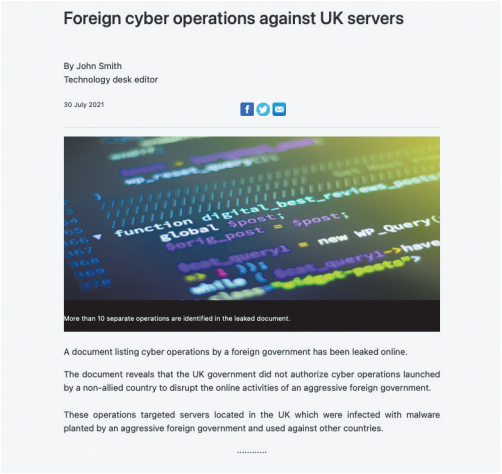
To test whether cyber operations produce a negative reaction among TPC publics (H_1) that is tempered by prior authorization (H_2) and existing alliances (H_3) and subject to the identity of the final target (H_4), the scenarios are constructed by manipulating the *authorization, alliance,* and *actor* treatments accordingly to create a scenario presented in a fictitious news article (see Figure 1). *Authorization* represents the operational context of the incident. The scenario depicts the operation as either authorized by the TPC beforehand or unsanctioned. If H_2 is valid, participants should express less negativity when told that this was an authorized operation. By contrast, both *alliance* and *actor* provide the strategic context. *Alliance* establishes whether an alliance exists between the initiator of the cyber operation and the TPC.[1] At the same time, *actor* reveals the final target's identity as either a state or non-state actor. If H_3 is valid, participants should express more negativity if an alliance does not exist. On the other hand, negativity is likely to increase regardless of prior authorization or alliance status if the final target is a state actor, as argued by H_4.

Readers should note that both the initiator and target of the cyber operation are intentionally unnamed. It is often the case that publicly disclosed information following a cybersecurity incident is ambiguous due to limited attributability or strategic necessity (Egloff and Smeets 2021). Moreover, identifying the source and final target of the operation may influence participants to base their decisions on established preconceptions towards individual state and non-state actors (Holsti 1967; Herrmann et al. 1997).

Depending on the treatment combination, participants may express higher or lower levels of negativity. Negativity is defined as the tendency to give more weight to negative (i.e., bad) information than to positive (i.e., good) information (Johnson and Tierney 2019). This is measured using the binary response variables *censure, continue,* and *consequences*.

1 The nature of the alliance is not explained further, as we believe that publics are unlikely to recognize the nuances surrounding this concept.

FIGURE 1: ARTICLE TREATMENT



Censure represents participant willingness to advocate policies aimed at penalizing the initiator. Participants who support penalization reflect increased negativity. Inversely, *continue* signals participant endorsement for continued cooperation with the initiator in future operations. Participants who support continued cooperation reflect decreased negativity. Finally, *consequences* indicate whether participants feel that the incident adversely affects TPC security. Participants who are concerned with the negative effects of the operation reflect increased negativity.

B. Covariates

The cybersecurity literature acknowledges competing explanations for the observed outcome. Specifically, established policy *preference* and cybersecurity *knowledge* may shape the public perception of incidents. Established policy preferences among participants may influence how they respond to the presented scenario. These preferences are captured using a modified version of the instrument developed by Maggiotto and Wittkopf (1981). The instrument measures support for militant, isolationist, and cooperative policies by instructing participants to evaluate their support for statements representing these policies using a seven-point Likert scale. *Preference* is a compound indicator[2] in which positive values indicate support for cooperative policies, while negative values suggest greater militancy. Values closer to zero represent isolationist tendencies.

Similarly, cybersecurity *knowledge* may also influence how participants respond to the scenario. This is measured using the instrument Gomez and Whyte (2021) developed, in which participants are asked a series of questions about cybersecurity concepts

2 Mean values reflecting support for militant, isolationist, and cooperative policies are computed as follows: (Cooperative – Militant) / Isolationist.

and incidents. Values closer to 1 indicate greater knowledge. Finally, demographic information is also collected.

To avoid the possibility of order effects,[3] the sequence of the experiment is randomized so that covariate information is collected either before or after the scenario and outcome variables. Furthermore, to validate the effects of the treatments, balance checks are performed to confirm successful randomization.

C. Recruitment

A priori power analysis establishes that 822 participants are necessary to achieve an alpha of 0.8 while conservatively expecting minimum treatment effects. Participants from the United Kingdom were selected given (1) the prevailing strategic environment and (2) existing alliance relationships. Tensions with Russia and the long-standing alliance with the United States suggest that the hypothesized effects are likely observed within this population.

To address naivete and participant engagement, common concerns surrounding Internet-based experiments, recruitment criteria, and attention checks were introduced.[4] Furthermore, only individuals who took part in previous studies and received reviews of 100% from other researchers on Prolific Academic were invited to participate. Finally, only individuals who had not taken part in the authors' previous experiments were eligible, to reduce the likelihood of familiarity with the experimental designs used.

5. RESULTS

The experiment recruited 899 participants. Their median age was 36.7 years (SD = 13.43), and they were almost evenly divided between male (50.4%) and female (49.6%) participants. Most (57.2%) report an annual income above the national median. In terms of education, 60.7% of participants possess a bachelor's degree or higher. Concerning the *preference* and *knowledge* covariates, the sample slightly prefers cooperative policies (0.5, SD = 0.48), while domain expertise is relatively low (0.45, SD = 0.27).

Regarding the *censure* variable, 600 (66.7%) participants, independent of treatment conditions, called for penalizing the initiator. Corresponding chi-square tests indicate that variations across treatment conditions are statistically significant. However, participant support for continued *cooperation* appears to be more evenly divided, with

3 The possibility that the order in which information is presented affects how the participants respond.

4 The popularity of Internet-based experiments means that participants can sometimes be aware of specific experimental treatments or are not engaged with the material. Attention checks and specific recruitment strategies address this concern.

415 (46.2%) expressing support irrespective of treatment conditions. As with *censure*, chi-square tests indicate that variations across treatment conditions are statistically significant. This, however, is not observed for the actor treatment. Finally, most participants (627, 69.7%) believe that the United Kingdom will suffer *consequences*. Corresponding chi-square tests indicate that variations across treatment conditions are statistically significant. These suggest that the *censure, cooperation,* and *consequences* variables correspond with the expectation that cybersecurity incidents, in general, increase negativity, as hypothesized by H_1.

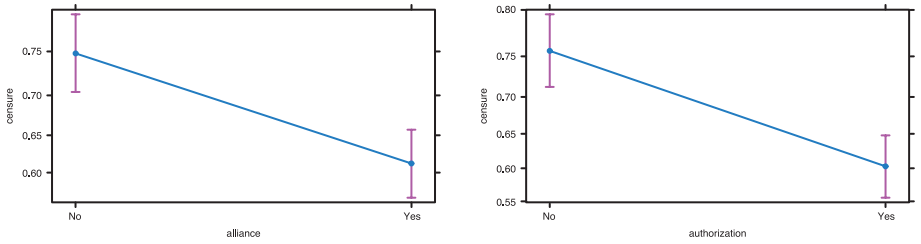
The specific causal effects of the *authorization, alliance,* and *actor* treatments on the *censure, cooperate,* and *consequence* outcome variables are depicted in the table shown in Figure 2. However, readers should note that the values reported in the text below are in log-odds for easier interpretation.

FIGURE 2: REGRESSION TABLE

	Dependent variable:		
	Censure	Cooperate	Consequences
Authorization	-0.718 (0.151)***	0.624 (0.142)***	-0.390 (0.153)*
Alliance	-0.633 (0.150)***	1.123 (0.142)***	-0.979 (0.156)***
Actor	0.287 (0.149)	-0.247 (0.142)	0.334 (0.153)*
Preference	0.067 (0.163)	-0.131 (0.154)	0.098 (0.169)
Knowledge	-1.446 (0.308)***	0.235 (0.290)	-0.294 (0.310)
Malware	-0.158 (0.462)	-0.296 (0.461)	1.188 (0.641)
Age	0.002 (0.006)	-0.002 (0.005)	-0.007 (0.006)
Male	-0.088 (0.169)	0.053 (0.160)	-0.480 (0.173)**
Education	0.035 (0.155)	0.115 (0.147)	-0.013 (0.159)
Income	0.061 (0.152)	-0.135 (0.145)	0.001 (0.156)
Constant	1.243 (0.297)***	-0.869 (0.282)**	1.921 (0.310)***
AIC	1089	1172.7	1048.9
McFadden	0.067	0.073	0.068
Note:			*p<0.05; **p<0.01; ***p<0.001

As indicated by the descriptive statistics, the operation appears to increase negativity among publics irrespective of prior authorization, existing alliances, and actor identity in support of H_1. However, public support for specific policies is influenced by these underlying conditions. *Censure* appears to be a function of both *authorization* and *alliance*. These decrease the likelihood of censuring the initiating actor by a factor of 0.487 and 0.531, respectively, and support the expectations of hypotheses H_2 and H_3. Simply put, the odds that TPC publics are likely to demand punishment if the operation is either sanctioned or conducted by an ally are approximately halved. Holding all other variables at their means, the effect of these two treatments is seen in Figure 3 below.

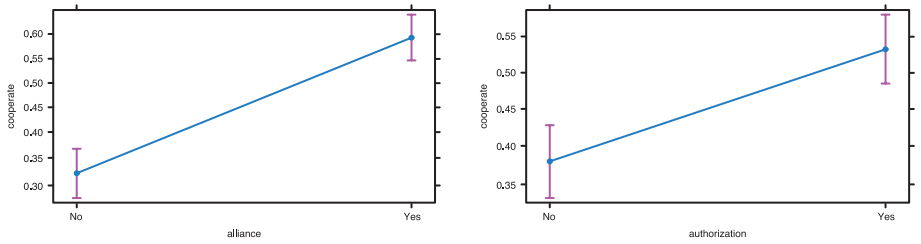
FIGURE 3: CENSURE EFFECTS PLOTS[5]



In keeping with the expectations of hypotheses H_2 and H_3, this observation is reversed for *cooperate*. Prior authorization and existing alliances increase the likelihood of continued cooperation by 1.866 and 3.074, respectively. Simply put, the odds that TPC publics would support continued cooperation are either approximately doubled if prior authorization exists or approximately tripled if an ally is involved. We see this effect in Figure 3 above, holding all other variables at their means.

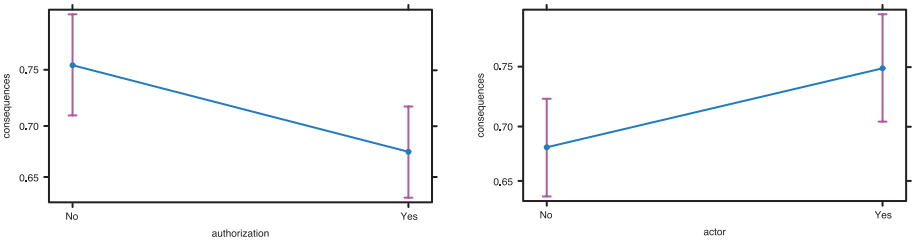
Finally, the results shown in the regression table (see Figure 2) indicate that the perceived consequences of a cyber operation appear to stem from all three treatments. Both *authorization* and *alliance* decrease concern surrounding the possible repercussion of a cyber operation by factors of 0.677 and 0.376, respectively, in keeping with the above hypotheses. Inversely, if the operation targets a state actor, public concern increases by a factor of 1.397. However, when all other variables are held at their means, the effects of both *authorization* and *actor* overlap with each other (see Figure 4), suggesting that alliance relationships are the primary driving force behind perceptions of risk in the context of this experiment, partially weakening the expectations of hypotheses H_2 and H_4. Simply put, the odds that TPC publics are concerned about possible consequences are nearly halved if an ally is involved.

FIGURE 4: COOPERATE EFFECTS PLOTS



5 For all successive effects plots, the data points represent the predicted probabilities for the respective treatments while holding other variables at their means (i.e., *censure, cooperate,* and *consequences*).

FIGURE 5: CONSEQUENCE EFFECTS PLOTS



While the treatments appear to influence *censure, cooperate,* and *consequence*, only *knowledge* and *male* covariates significantly influence the outcome. Greater domain expertise tempers support for censuring by a factor of 0.236, suggesting that these individuals either are aware that cyber operations are increasingly a part of modern interstate interactions (Gomez and Whyte 2021) or are uncertain of the utility afforded by this course of action (Kreps and Schneider 2019; Schneider 2017). This finding is not surprising if one assumes that publics, in general, recognize cyber operations as a feature of modern interstate interactions. If this is the case, participants may hesitate to censure an initiator without further strategic context.

Interestingly, male participants appear less concerned with the potential consequences of a cyber operation and reduce the likelihood of this outcome by a factor of 0.619. This is an interesting finding, as gender does not shape threat perception towards cyberspace in other research (Gomez 2019a; Gomez 2019b; Kreps and Schneider 2019; Shandler et al. 2021). Although research involving conventional conflict suggests that gender may influence risk tolerance (e.g., casualty aversion), this topic requires further investigation beyond the scope of this article (Crawford, Lawrence, and Lebovic 2017).

6. CONCLUSION

With states employing cyber operations as an instrument of foreign policy and considering the latent interconnectivity characterizing cyberspace, the involvement of TPCs is an increasingly common and almost inevitable feature of cyber conflict. We find that the inclusion of TPCs carries political risks for perpetrators, and operational planners cannot ignore the risk of widespread backlash from a TPC population. Whereas the existing literature finds that public opinion is inflamed by cyberattacks (Kreps and Schneider 2019; Gomez and Whyte 2021; Shandler et al. 2021a), our project shows that this is not merely a question of victimization. Instead, our study of the United Kingdom indicates that publics believe that national sovereignty extends

into cyberspace and care about what happens in their national networks even if they are not being directly targeted. That this ire was successfully assuaged by the presence of an alliance or prior approval suggests that while coordinating with a TPC government may be onerous, it can have meaningful benefits for operational planners and certainly reduce the potential for future fallout if an operation becomes public knowledge.

From a tactical perspective, prior authorization to enable future operations while avoiding cost imposition by TPCs imposes constraints on speed, intensity, and control, contributing to the trilemma first proposed by Maschmeyer (2021). Coordination lengthens the time necessary to deploy cyber operations and provides targets with the ability to introduce additional security measures. Similarly, TPCs concerned with possible second- or third-order consequences may impose constraints on the scope of the operation or even the techniques employed. Indeed, as with the recent "hunt forward" missions, control may need to be shared to consider TPC interests, further constraining the initiator. Although compromises are possible, TPC involvement introduces restrictions that may impact strategic objectives. Our findings suggest that while such back-and-forth between an attack-initiator and a TPC government may be burdensome, it may prove essential for avoiding blowback and subsequent fallout.

Furthermore, this project shows the continued influence of established relationships on popular preferences and perceptions (Tomz and Weeks 2021) and the implications of such dynamics for more active cyber strategies. Take the case of the US approach towards persistent engagement and the use of externally oriented operations as a means of advancing strategic objectives. While "defend forward" is depicted as the preferred means of protecting US capabilities and interests, it does not consider public opinion or how the nature of relationships (i.e., allies versus partners versus adversaries) influence this strategic preference. Repeated interaction in cyberspace may lead to stability (Fischerkeller and Harknett 2018), but such events do not occur in a vacuum and cannot be achieved while sacrificing the diplomatic and political capital in TPCs needed to sustain such outcomes. That alliances and prior approval temper such widespread backlash indicates that offensive cyber operations need not be abandoned as a tool of statecraft. However, to be genuinely effective at achieving sustained progress towards stability, such endeavors also require persistence in proactive diplomatic measures and relationship-building with TPCs to sustain international support and avoid sub-optimal policy choices. Indeed, just as Status of Forces Agreements have emerged to govern conventional military means, similar diplomatic and legal measures in cyberspace can potentially resolve the current uncertainty and provide a firm guide for operational planners.

The introduction of cyber operations as an additional instrument in states' foreign policy toolboxes broadens policy options for those able to exploit this environment.

Scholars and policymakers alike, however, need to recognize that actions in cyberspace are not shaped solely by the underlying technologies. The intentional (or unintentional) involvement of TPCs in these operations and the consequences that may emerge from their inclusion cannot be ignored and need to be contained within the strategic calculus of the digital domain.

REFERENCES

Acharya, Amitav. 2007. "State Sovereignty after 9/11: Disorganised Hypocrisy." *Political Studies* 55, no. 2: 274–296.

Biswas, Bidisha. 2009. "Just Between Friends: Bilateral Cooperation and Bounded Sovereignty in the 'Global War on Terror.'" *Politics and Policy* 37, no. 5: 929–950.

Buhaug, Halvard, and Kristian Skrede Gleditsch. 2008. "Contagion or Confusion? Why Conflicts Cluster in Space." *International Studies Quarterly* 52, no. 2: 215–233.

Clarke, Richard A. 2004. *Against All Enemies: Inside America's War on Terror*. Simon and Schuster.

Craig, Anthony. 2020. "Capabilities and Conflict in the Cyber Domain: An Empirical Study." PhD dissertation, Cardiff University.

Crawford, Kerry F., Eric D. Lawrence, and James H. Lebovic. 2017. "Aversion, Acceptance, or Apprehension? The Effects of Gender on US Public Opinion Concerning US-Inflicted Civilian Casualties." *Journal of Global Security Studies* 2, no. 2: 150–169.

Cronin, Audrey Kurth. 2002. "Rethinking Sovereignty: American Strategy in the Age of Terrorism." *Survival* 44, no. 2: 119–139.

Department of Defense. 2016. "Agreed Operation Glowing Symphony Notification Plan." National Security Archives, November 4. https://nsarchive.gwu.edu/.

DeSombre, Winnona, Michele Campobasso, Luca Alodi, James Shires, J. D. Work, Robert Morgus, Patrick Howell O'Neill, and Trey Herr. 2021. "A Primer on the Proliferation of Offensive Cyber Capabilities," March 1. https://www.atlanticcouncil.org/in-depth-research-reports/issue-brief/a-primer-on-the-proliferation-of-offensive-cyber-capabilities/.

Dunn Cavelty, Myriam. 2013. "From Cyber-Bombs to Political Fallout: Threat Representations with an Impact in the Cyber-Security Discourse." *International Studies Review* 15, no. 1: 105–122.

Easley, Leif-Eric. 2014. "Spying on Allies." *Survival* 56, no. 4: 141–156.

Egloff, Florian J., and Max Smeets. 2021. "Publicly Attributing Cyber Attacks: A Framework." *Journal of Strategic Studies*, 1–32.

Enders, Walter, and Todd Sandler. 2002. "Patterns of Transnational Terrorism, 1970–1999: Alternative Time-Series Estimates." *International Studies Quarterly* 46, no. 2: 145–165.

Fischerkeller, Michael P., and Richard J. Harknett. 2018. "Persistent Engagement, Agreed Competition, Cyberspace Interaction Dynamics and Escalation." *Orbis* 61, no. 3: 381–393.

Fischerkeller, Michael P., and Richard J. Harknett. 2020. "Cyber Persistence, Intelligence Contests, and Strategic Competition." Part 5 of "Policy Roundtable: Cyber Conflict as an Intelligence Contest," in "Cyber Competition," special issue, *Texas National Security Review*, September 17. https://tnsr.org/roundtable/policy-roundtable-cyber-conflict-as-an-intelligence-contest/.

Gazis, Olivia. 2021. “US Launched ‘More Than 2 Dozen’ Cyber Operations to Protect Election.” CBS News, March 25. https://www.cbsnews.com/news/election-interference-us-cyber-command-nsa-nakasone/.

Gigerenzer, Gerd. 2006. “Out of the Frying Pan into the Fire: Behavioral Reactions to Terrorist Attacks.” *Risk Analysis: An International Journal* 26, no. 2: 347–351.

Gleditsch, Kristian Skrede, Idean Salehyan, and Kenneth Schultz. 2008. “Fighting at Home, Fighting Abroad: How Civil Wars Lead to International Disputes.” *Journal of Conflict Resolution* 52, no. 4: 479–506.

Gomez, Miguel Alberto. 2019a. “Past Behavior and Future Judgements: Seizing and Freezing in Response to Cyber Operations.” *Journal of Cybersecurity* 5, no. 1: tyz012.

Gomez, Miguel Alberto. 2019b. “Sound the Alarm! Updating Beliefs and Degradative Cyber Operations.” *European Journal of International Security* 4, no. 2: 190–208.

Gomez, Miguel Alberto, and Eula Bianca Villar. 2018. “Fear, Uncertainty, and Dread: Cognitive Heuristics and Cyber Threats.” *Politics and Governance* 6, no. 2: 61–72.

Gomez, Miguel Alberto, and Christopher Whyte. 2021. “Breaking the Myth of Cyber Doom: Securitization and Normalization of Novel Threats.” *International Studies Quarterly*.

Gross, Michael L., Daphna Canetti, and Dana R. Vashdi. 2017. “Cyberterrorism: Its Effects on Psychological Well-Being, Public Confidence and Political Attitudes.” *Journal of Cybersecurity* 3, no. 1: 49–58.

Hansen, Lene, and Helen Nissenbaum. 2009. “Digital Disaster, Cyber Security, and the Copenhagen School.” *International Studies Quarterly* 53, no. 4: 1155–1175.

Herrmann, Richard K., James F. Voss, Tonya Y. E. Schooler, and Joseph Ciarrochi. 1997. “Images in International Relations: An Experimental Test of Cognitive Schemata.” *International Studies Quarterly* 41, no. 3: 403–433.

Holsti, Ole R. 1967. “Cognitive Dynamics and Images of the Enemy.” *Journal of International Affairs* 21, no. 1: 16–39.

Hughes, Geraint. 2014. “Skyjackers, Jackals and Soldiers: British Planning for International Terrorist Incidents during the 1970s.” *International Affairs* 90, no. 5: 1013–1031.

Jackson, Robert. 2007. “Sovereignty and Its Presuppositions: Before 9/11 and After.” *Political Studies* 55, no. 2: 297–317.

Jarvis, Lee, Stuart Macdonald, and Andrew Whiting. 2017. “Unpacking Cyberterrorism Discourse: Specificity, Status, and Scale in News Media Constructions of Threat.” *European Journal of International Security* 2, no. 1: 64–87.

Johnson, Dominic D. P., and Dominic Tierney. 2018. “Bad World: The Negativity Bias in International Politics.” *International Security* 43, no. 3: 96–140.

Kertzer, Joshua D., and Thomas Zeitzoff. 2017. “A Bottom-Up Theory of Public Opinion about Foreign Policy.” *American Journal of Political Science* 61, no. 3: 543–558.

Klein, Aaron J. 2007. *Striking Back: The 1972 Munich Olympics Massacre and Israel’s Deadly Response*. Random House.

Kostyuk, Nadiya, and Carly Wayne. 2021. “The Microfoundations of State Cybersecurity: Cyber Risk Perceptions and the Mass Public.” *Journal of Global Security Studies* 6, no. 2: ogz077.

Kreps, Sarah. 2010. “Elite Consensus as a Determinant of Alliance Cohesion: Why Public Opinion Hardly Matters for NATO-Led Operations in Afghanistan.” *Foreign Policy Analysis* 6, no. 3: 191–215.

Kreps, Sarah, and Jacquelyn Schneider. 2019. “Escalation Firebreaks in the Cyber, Conventional, and Nuclear Domains: Moving beyond Effects-Based Logics.” *Journal of Cybersecurity* 5, no. 1: tyz007.

Lawson, Sean, and Michael K. Middleton. 2019. “Cyber Pearl Harbor: Analogy, Fear, and the Framing of Cyber Security Threats in the United States, 1991–2016.” *First Monday* 24, no. 3 (March 4, 2019).

Libicki, Martin C. 2009. *Cyberdeterrence and Cyberwar*. RAND Corporation.

Lin, H. (2012). Cyber conflict and international humanitarian law. *International Review of the Red Cross*, 94(886), 515–531.

Maggiotto, Michael A., and Eugene R. Wittkopf. 1981. “American Public Attitudes toward Foreign Policy.” *International Studies Quarterly* 25, no. 4: 601–631.

Maschmeyer, Lennart. 2021. “The Subversive Trilemma: Why Cyber Operations Fall Short of Expectations.” *International Security* 46, no. 2: 51–90.

Mason, R. C. 2010. *Status of Forces Agreement: What Is It, and How Has It Been Utilized?* DIANE Publishing.

Nakashima, Ellen, 2017. “US Military Cyber Operation to Attack ISIS Last Year Sparked Heated Debate over Alerting Allies.” *Washington Post*, May 9.

Nance, Malcolm, and Christopher Sampson. 2017. *Hacking ISIS: How to Destroy the Cyber Jihad*. Simon and Schuster.

Otto, Jacob, and William Spaniel. 2021. “Doubling Down: The Danger of Disclosing Secret Action.” *International Studies Quarterly* 65, no. 2: 500–511.

Owen, M., and Maurer, K. 2014. *No Easy Day: The Firsthand Account of the Mission that Killed Osama Bin Laden*. Penguin.

Patman, Robert G. 2006. “Globalisation, the New US Exceptionalism and the War on Terror.” *Third World Quarterly* 27, no. 6: 963–986.

Pettyjohn, S. L., and J. Kavanagh. 2016. *Access Granted: Political Challenges to the US Overseas Military Presence, 1945–2014*. RAND Corporation.

Pop, Valentina, and Sune Engel Rasmussen. 2018. “Islamic State Propaganda Sites Shut Down.” *Wall Street Journal*, April 27.

Putnam, Robert D. 1988. “Diplomacy and Domestic Politics: The Logic of Two-Level Games.” *International Organization* 42, no. 3: 427–460.

Rapport, David (2004). “The Four Waves of Modern Terrorism.” In *Attacking Terrorism: Elements of a Grand Strategy*. Audrey Kurth Cronin and James M. Ludes (Eds.), Washington, DC: Georgetown University Press, 46–73.

Reeve, Simon. 2000. *One Day in September: The Full Story of the 1972 Munich Olympics Massacre and the Israeli Revenge Operation “Wrath of God.”* Skyhorse Publishing.

Reinhardt, Gina Yannitell. 2017. “Imagining Worse Than Reality: Comparing Beliefs and Intentions between Disaster Evacuees and Survey Respondents.” *Journal of Risk Research* 20, no. 2: 169–194.

Rovner, Joshua. 2020. “What is an Intelligence Contest?” Part 2 of “Policy Roundtable: Cyber Conflict as an Intelligence Contest,” in “Cyber Competition,” special issue, *Texas National Security Review*, September 17. https://tnsr.org/roundtable/policy-roundtable-cyber-conflict-as-an-intelligence-contest/.

Schaller, Christian. 2015. “Using Force against Terrorists ‘Outside Areas of Active Hostilities’: The Obama Approach and the Bin Laden Raid Revisited.” *Journal of Conflict and Security Law* 20, no. 2: 195–227.

Schmitt, Michael N., ed. 2017. *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*. Cambridge University Press.

Schmitt, Eric, and Thom Shanker. 2011. *Counterstrike: The Untold Story of America's Secret Campaign against al Qaeda*. Macmillan.

Schneider, Jacquelyn. 2017. "Cyber and Crisis Escalation: Insights from Wargaming." *USASOC Futures Forum*.

Shandler, Ryan, Michael L. Gross, Sophia Backhaus, and Daphna Canetti. 2021a. "Cyber Terrorism and Public Support for Retaliation: A Multi-Country Survey Experiment." *British Journal of Political Science*, 1–19.

Shandler, Ryan, Michael L. Gross, and Daphna Canetti. 2021b. "A Fragile Public Preference for Cyber Strikes: Evidence from Survey Experiments in the United States, United Kingdom, and Israel." *Contemporary Security Policy* 42, no. 2: 135–162.

Siverson, Randolph M., and Harvey Starr. 1991. *The Diffusion of War: A Study of Opportunity and Willingness*. University of Michigan Press.

Smith, Gregory L. 2019. "Secret but Constrained: The Impact of Elite Opposition on Covert Operations." *International Organization* 73, no. 3: 685–707.

Temple-Raston, Dina. 2019. "How the US Hacked ISIS." NPR [National Public Radio], September 26. https://www.npr.org/2019/09/26/763545811/how-the-u-s-hacked-isis.

Tomz, Michael, and Jessica L. P. Weeks. 2021. "Military Alliances and Public Support for War." *International Studies Quarterly* 65, no. 3: 811–824.

Vavra, Shannon. 2020. "Cyber Command Deploys Abroad to Fend off Foreign Hacking ahead of the 2020 Election." August 25. https://www.cyberscoop.com/2020-presidential-election-cyber-command-nakasone-deployed-protect-interference-hacking/.

Viscusi, W. Kip, and Richard J. Zeckhauser. 2017. "Recollection Bias and Its Underpinnings: Lessons from Terrorism Risk Assessments." *Risk Analysis* 37, no. 5: 969–981.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

Machine Expertise in the Loop: Artificial Intelligence Decision-Making Inputs and Cyber Conflict

Christopher E. Whyte
L. Douglas Wilder School of
Government and Public Affairs
Virginia Commonwealth University

Abstract: How will national cybersecurity stakeholders react to increasingly sophisticated artificial intelligence (AI) products in the decision-making loop? Diverse AI applications, already a component of military operations, stand to further revolutionize threat-intelligence analytics, curation, and presentation in years to come. In doing so, they portend heightened efficiency and higher tempos of security operations while minimizing risk potential. And yet, challenges abound. AI inputs to deliberative processes may be perceived as more or less robust, accurate, or generalizable according to a range of factors that influence decision-makers. Then these habits and preferences might be reintroduced to the information loop in several ways, as the algorithm, the design/implementation process, and institutions adapt to accommodate the human element. Malicious actors are even likely to target this action-reaction loop to influence AI systems. This two-phase cycle is perhaps most worrisome in relation to cyber conflict, where informational ambiguity, functionally diverse workforces, and an expansive and fragmented threat environment combine to produce an immense opportunity for baking bias into the loop. This paper presents the results of two experiments designed to explore different manifestations of AI systems in the cyber conflict decision-making loop. Though findings suggest that technical expertise positively impacts respondents' ability to gauge the potential utility and credibility of an input (indicating that training can, in fact, overcome decision-maker bias), the perception of human agency in the loop even in the presence of AI inputs mitigates this cautionary effect and makes decision-makers more willing to operate on less overall information.

Keywords: *artificial intelligence, decision-making, cyber operations*

1. INTRODUCTION

The impact of AI systems[1] and machine learning on the shape, scope, and aims of sophisticated cyber operations is a question of great interest among scholars[2] and practitioners focused on issues of global cyber conflict.[3] The inclusion of mechanisms for rapid analysis, intrusion, and reconfiguration stands to both introduce immense new opportunities for the novel compromise of target systems and enhance existing methods thereof. Likewise, Big Data analytic tools stand to increase the command-and-control capacities of those seeking to wield cyber operations for strategic gains, further enhancing capabilities. Naturally, defenders also stand to benefit from AI.[4] But it also seems increasingly likely that a singular focus on the offense-defense balance in cyberspace in the context of AI—or, at least, on its purely technological components—may mislead many as to the novel risks associated with increased machine expertise in the digital conflict loop.[5] After all, expanded use of AI across national security establishments adds new incentives to interfere with machine systems, an activity that can primarily be accomplished via cyberspace.[6] This suggests an expansion of the reasons why defenders—particularly government and government-affiliated defenders—may encounter sophisticated offensive cyber activities over time beyond just evolving technical capabilities. Moreover, the more that novel AI mechanics are included in such activities, the more the perception of decision-makers of adversary intentions and presence becomes an unpredictable element in any given cyber episode.

That the human side of using AI for cyber operations must be considered and prioritized amid efforts to operationalize technological breakthroughs is clearly known to security practitioners. In late 2021, a high-ranking officer in the United States Marine Corps, Lt. Gen. Michael S. Groen, stated that the organization was beginning to undergo a necessary "mind shift" on the role of network architecture in the age of non-human intelligence and that "a culture shift [would] be needed in

1 For some descriptions, see, e.g., Thanh Cong Truong, Quoc Bao Diep, and Ivan Zelinka, "Artificial Intelligence in the Cyber Domain: Offense and Defense," *Symmetry* 12, no. 3 (2020): 410; and Adrien Bécue, Isabel Praça, and João Gama, "Artificial Intelligence, Cyber-Threats and Industry 4.0: Challenges and Opportunities," *Artificial Intelligence Review* 54, no. 5 (2021): 3849–3886.

2 Benjamin M. Jensen, Christopher Whyte, and Scott Cuomo, "Algorithms at War: The Promise, Peril, and Limits of Artificial Intelligence," *International Studies Review* 22, no. 3 (2020): 526–550.

3 Prominently, e.g., Mariarosaria Taddeo and Luciano Floridi, "Regulate Artificial Intelligence to Avert Cyber Arms Race," Nature 556 (2018): 296–298; Amandeep Singh Gill, "Artificial Intelligence and International Security: The Long View," *Ethics and International Affairs* 33, no. 2 (2019): 169–179.

4 Enn Tyugu, "Artificial Intelligence in Cyber Defense," in *2011 3rd International Conference on Cyber Conflict* (IEEE, 2011), 1–11; Caitriona H. Heinl, "Artificial (Intelligent) Agents and Active Cyber Defence: Policy Implications," in *2014 6th International Conference On Cyber Conflict (CyCon 2014)* (IEEE, 2014), 53–66.

5 James Johnson, "The AI-Cyber Nexus: Implications for Military Escalation, Deterrence and Strategic Stability," *Journal of Cyber Policy* 4, no. 3 (2019): 442–460.

6 Christopher Whyte, "Problems of Poison: New Paradigms and 'Agreed' Competition in the Era of AI-Enabled Cyber Operations," in *2020 12th International Conference on Cyber Conflict (CyCon)*, vol. 1300 (IEEE, 2020), 215–232.

order for operators and warfighters to embrace AI integration."[7] Without operators and commanding officers who understand the limitations of machine learning and intelligent agents employed to streamline cyber operations, security institutions risk many potential pitfalls through their use. These include a simple inability to harness the intended benefits of using AI, as well as tactical overstep and real potential conflict escalation stemming from a misreading of what AI really adds to cyber conflict.[8]

Few empirical baselines for understanding the impact of AI systems introduced to conflict processes exist at present. Moreover, thus far there is limited evidence upon which to base an assessment of what might prime decision-makers towards different kinds of behavioral outcomes as they encounter AI in real-world conflict scenarios. Indeed, the scope of cyber-psychological studies of decision-making itself remains limited at the time of writing, even given the abundance of recent scholarly publications on the topic.[9] As a result, the best approach to initially problematizing the challenge of AI for cyber conflict decision-makers is direct evidence gathering.

Following a summary of contemporary thinking on the role of AI in global cyber conflict and the uncertainties it may bring, the remainder of this article describes such a data-gathering effort, involving two surveys of separate groups of several hundred national security practitioners exposed to different cyber conflict scenarios involving AI. I find immense differentiation in the manner decision-makers react to AI based on its integration as either a tactical or a strategic tool. When encountered in the decision-making loop as a holistic presence shaping foreign policy events, elite respondents demonstrate restraint, primarily as a result of fear about the error-proneness of AI systems. However, this caution is mitigated in situations where it appears as though a human presence remains in the loop, a potentially worrying dynamic given that control groups in the studies were more assertive in responding to perceived cyber aggression.

7 David Vergun, "General Says Artificial Intelligence Will Play Important Role in Network Defense," U.S. Department of Defense, 8 October 2021, https://www.defense.gov/News/News-Stories/Article/Article/2805760/general-says-artificial-intelligence-will-play-important-role-in-network-defense/.

8 James Johnson, "Artificial Intelligence, Drone Swarming and Escalation Risks in Future Warfare," *RUSI Journal* 165, no. 2 (2020): 26–36; James Johnson, "Catalytic Nuclear War in the Age of Artificial Intelligence and Autonomy: Emerging Military Technology and Escalation Risk between Nuclear-Armed States," *Journal of Strategic Studies* (2021): 1–41.

9 E.g., Ryan Shandler, Michael L. Gross, Sophia Backhaus, and Daphna Canetti, "Cyber Terrorism and Public Support for Retaliation: A Multi-Country Survey Experiment," *British Journal of Political Science* (2021): 1–19; Miguel Alberto Gomez and Christopher Whyte, "Breaking the Myth of Cyber Doom: Securitization and Normalization of Novel Threats," *International Studies Quarterly* 65, no. 4 (2021): 1137–1150.

2. ARTIFICIAL INTELLIGENCE AND CYBER CONFLICT

Artificial intelligence is a catchall term that encompasses an array of technological and scientific advances emerging from an array of disciplines.[10, 11] Recent work on cyber conflict has highlighted AI as a prospective game-changer when it comes to the planning and execution of cyber operations, as well as for efforts to defend against and deter cyber aggression.[12] However, as empirical references remain limited and largely hypothetical—at least in non-classified settings—so too does the opportunity for extrapolating from technical developments to likely operational and strategic effects. At the time of writing, it would not be unreasonable to characterize most assessments of AI's potential impact on cybersecurity as a move towards more of the same, just "faster, smarter, bigger, better." Specifically, experts appear to fear that AI will extend the lifecycle of cyber threats and create an expanded footprint for malicious code that will be more difficult for defenders to model, detect, and mitigate.[13] Likewise, machine learning stands to make sophisticated malicious cyber action more accessible to the average security actor, a dynamic that further enlarges the toolkit of tricks that attackers must rely upon and defenders must consider.[14]

It is not unreasonable to consider potential transformations as occurring across three oft-cited levels of analysis—the (1) tactical, (2) operational, and (3) strategic levels of security operation. At the tactical level, wherein direct inter-adversary engagement is planned and executed, the increasing opportunity to design software that is highly adaptive beyond the requirement for human input to the process portends substantial challenges for cyber defense.[15] Three categories of such adaptability deserve particular mention. First, AI-augmented malware implies an emergent capability for technique selection.[16] An autonomous ability to assess environments and select the optimal method of intrusion is worrisome, not so much because it differs dramatically from existing capabilities among sophisticated threat actors but because a more accessible version of the capacity implies a potential rise in the average sophistication of cyber threats. Second, and on a related note, advanced malware might prove itself capable of tactical adaptation beyond simple method selection (i.e., an ability to abandon one

10 Benjamin M. Jensen, Christopher Whyte, and Scott Cuomo, "Algorithms at War: The Promise, Peril, and Limits of Artificial Intelligence," *International Studies Review* 22, no. 3 (2020): 526–550.

11 For robust histories of the development of the AI field, see, e.g., Nils J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements* (New York: Cambridge University Press, 2010); Herbert Simon, "Artificial Intelligence: An Empirical Science," *Artificial Intelligence* 77, no. 2 (1995): 95–127.

12 See, e.g., Christopher Whyte, "Poison, Persistence, and Cascade Effects," *Strategic Studies Quarterly* 14, no. 4 (2020): 18–46; Joe Burton and Simona R. Soare, "Understanding the Strategic Implications of the Weaponization of Artificial Intelligence," in *2019 11th International Conference on Cyber Conflict (CyCon)*, vol. 900 (IEEE, 2019), 1–17; Zachary Davis, "Artificial Intelligence on the Battlefield," *Prism* 8, no. 2 (2019): 114–131.

13 Whyte, "Poison, Persistence, and Cascade Effects."

14 Ibid.; Pavel Sharikov, "Artificial Intelligence, Cyberattack, and Nuclear Weapons: A Dangerous Combination," *Bulletin of the Atomic Scientists* 74, no. 6 (2018): 368–373.

15 Whyte, "Poison, Persistence, and Cascade Effects."

16 Ibid., 24.

tactic in favor of another based on analysis of environment contours and defenses).[17] Finally, AI may lend itself not only to a technical ability to select tactical orientation but also to an ever-increasing capacity for value calculations vis-à-vis strategic objectives.[18] An AI-enabled ability to "use incoming data obtained via infection of machines to probabilistically judge where and when further infection is likely to lead to some value return"[19] offers a tactical salve for the constraint of necessary operational resource commitment that often plagues offensive cyber operations (OCO) planning.[20]

Operationally, machine learning and Big Data analytics most clearly imply an expanding capability for intelligence agencies, security institutions, and criminal enterprises alike to develop extremely high-fidelity comprehension of the attack surface of targets, whether those be individuals, facilities, institutions, or even national and sub-national units.[21] Simply put, the more information made available to competent cybersecurity stakeholders, the more capable the infrastructure and tools developed for malicious cyber activity are likely to be.[22] For defenders, this operational advantage suggests heightened insecurity on two particular fronts. First, it implies a general diversification in the tools available for procurement by either criminal or political security actors in the artifact- and cyber-threat-services development marketplaces.[23] Second, and more directly, it implies an ever-increasing probability of compromise via lateral access to critical systems (i.e., access gained via indirect targeting of associated institutions, users, or architecture).[24]

It is at the strategic level where the impact of varied AI technologies and advances on the contours of global cyber conflict remains the most difficult to forecast. On the one hand, everything discussed above implies an increased tempo of engagement between capable cyber actors in world affairs, as well as an expansion of the landscape of those that might qualify as "capable" and a pressing need to rely on machine solutions for challenges whose scale and scope might increasingly be shaped more by algorithmic input than human design.[25] On the other hand, such a read of the future of AI and

17 Ibid., 24–25.
18 Ibid.
19 Ibid., 24.
20 Christopher Whyte and Brian Mazanec, *Understanding Cyber Warfare: Politics, Policy and Strategy* (Routledge, 2018), chap. 5.
21 Louise Leenen and Thomas Meyer, "Artificial Intelligence and Big Data Analytics in Support of Cyber Defense," in *Research Anthology on Artificial Intelligence Applications in Security* (IGI Global, 2021), 1738–1753. See also Hamid Jahankhani, Stefan Kendzierskyj, Nishan Chelvachandran, and Jaime Ibarra, eds., *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity* (Springer Nature, 2020).
22 Whyte, "Poison, Persistence, and Cascade Effects," 25–26.
23 Murat Kantarcioglu and Fahad Shaon, "Securing Big Data in the Age of AI," in *2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)* (IEEE, 2019), 218–220.
24 Petar Radanliev, David De Roure, Rob Walton, Max Van Kleek, Rafael Mantilla Montalvo, La'Treall Maddox, Omar Santos, Peter Burnap, and Eirini Anthi, "Artificial Intelligence and Machine Learning in Dynamic Cyber Risk Analytics at the Edge," *SN Applied Sciences* 2, no. 11 (2020): 1–8.
25 Whyte, "Poison, Persistence, and Cascade Effects."

cyber conflict (i.e., the "bigger, faster, smarter" paradigm) ignores the reality that AI incorporated across all facets of national security and society writ large portends new rationales for cyber engagement.[26]

Simply put, as AI systems themselves become synonymous with critical societal and security functions, they become targets of extreme value. Specifically, a malicious actor might either (1) target AI systems with *input attacks* that present intentionally misleading observations to a machine learning algorithm so as to guide its calculations or (2) engage in *poisoning* activities that actually corrupt a learner's dataset, algorithm, or models via the provision of manipulated data.[27] In doing so, they might subvert the functionality of systems both bound up in the workings of the fifth domain and beyond it. And while there are, of course, a great many ways to undertake such interference, cyberspace is arguably the most direct and critical avenue by which to do so. Thus the strategic implication of AI is not only that current trends are intensifying but also that the subversive character of cyberspace is set to take on new dimensionality as cyber-enabled artificial intelligence attacks (CAIA) become an appealing method for interference, espionage, crime, and sabotage even beyond the scope of domain-specific activities.[28]

3. A STUDY OF MACHINES IN THE CYBER CONFLICT DECISION-MAKING LOOP

Past research informs us that leaders, military strategists, and technical practitioners alike are motivated, variously, by knowledge of the context of foreign policy situations,[29] by an in-depth technical understanding of the operational realities of state security operations,[30] by institutional imperatives,[31] and by personal affective and cognitive effects.[32] Artificial intelligence adds a new dimension to foreign-policy-crisis politics.[33] This is particularly the case for cyber conflict and "gray zone"

26 Michael C. Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power," *Texas National Security Review* 1, no. 3 (May 2018): 36–57.

27 Marcus Comiter, "Attacking Artificial Intelligence," Belfer Center Paper, Belfer Center, 2019. See also David Rios Insua, Roi Naveiro, Victor Gallego, and Jason Poulos, "Adversarial Machine Learning: Perspectives from Adversarial Risk Analysis," https://doi.org/10.48550/arXiv.2003.03546 (2020); Deepak Puthal and Saraju P. Mohanty, "Cybersecurity Issues in AI," *IEEE Consumer Electronics Magazine* 10, no. 4 (2021): 33–35.

28 Whyte, "Poison, Persistence, and Cascade Effects," 26.

29 Bruce Bueno De Mesquita, "Toward a Scientific Understanding of International Conflict: A Personal View," *International Studies Quarterly* 29, no. 2 (1985): 121–136.

30 Robert L. Pfaltzgraff, Jr., *Contending Theories of International Relations: A Comprehensive Survey* (Longman, 1997).

31 Stephen Van Evera, "The Cult of the Offensive and the Origins of the First World War," *International Security* 9, no. 1 (1984): 58–107.

32 Jonathan Mercer, "Emotional Beliefs," *International Organization* 64, no. 1 (2010): 1–31; Rose McDermott, "The Feeling of Rationality: The Meaning of Neuroscientific Advances for Political Science," *Perspectives on Politics* 2, no. 4 (2004): 691–706.

33 Margaret G. Hermann, Thomas Preston, Baghat Korany, and Timothy M. Shaw, "Who Leads Matters: The Effects of Powerful Individuals," *International Studies Review* 3, no. 2 (2001): 83–131.

operations in which peer competitors leverage sub-optimal mechanisms of state power—such as cyber operations, the deployment of unmarked soldiers to conflict zones, and the use of merchant vessels to harass shipping—to avoid escalation.[34] The stakes of producing better knowledge about the prospective effects of AI on a variety of cyber conflict dynamics are high. However, thus far there has been no specific attempt to examine how elite populations react to the input of AI during cyber conflict episodes, a notable shortcoming given the strength of voices across Western defense establishments now calling for changes to culture and practice to best harness new AI potential for national security.

A. Methods

The empirical study presented here considers the impact of AI on decision-making during cyber conflict episodes in two distinct dimensions. First, I consider the degree to which variable levels of AI-generated intelligence about conflict activities provided to elite decision-makers are likely to influence their support for different measures of preemptive action aimed at shutting down a prospective attack by an assertive foreign state actor and deterring future aggression. Second, I consider several dimensions of AI use by adversaries as they work on a sophisticated cyber campaign centered on political interference.

These paired studies take the form of small-scale survey experiments that engage populations drawn from several professional military education (PME) institutions in the United States and Central Europe.[35] Study #1 was run with 286 mid-career military professionals and faculty at two institutions in the United States and Germany. Study #2 was run with 503 such individuals, alongside invited workshop participants, at three institutions in the United States, Germany, and Slovenia. All exercises were conducted between June 2020 and August 2021 via an online interface. Pre-study questionnaires captured a full range of demographic control information to aid analysis, including occupation, nationality, age, education, and, following well-established standards in the international relations (IR) field,[36] foreign policy outlook and subject-specific know-how (both reported as an index value drawn from several questions). Follow-on debriefing was conducted via Zoom sessions for Study #1 and as a series of open-ended feedback prompts with Study #2.

Study #1 consists of a relatively straightforward experimental design based around a scenario of conflictual cyber activities. Participants were presented with a developing

34 Tim Stevens, "Knowledge in the Grey Zone: AI and Cybersecurity," *Digital War* 1, no. 1 (2020): 164–170.

35 Survey experiments are a common tool of IR research, and war-game-style experimental designs are increasingly being employed, particularly for situations where elite populations may differ in their approach vis-à-vis general populations and where there exists unusual uncertainty, such as with cyber conflict. See Erik Lin-Greenberg, Reid B. C. Pauly, and Jacquelyn G. Schneider, "Wargaming for International Relations Research," *European Journal of International Relations* (2021), https://doi.org/10.1177/13540661211064090.

36 E.g., in Joshua D. Kertzer, Kathleen E. Powers, Brian C. Rathbun, and Ravi Iyer, "Moral Support: How Moral Values Shape Foreign Policy Attitudes," *Journal of Politics* 76, no. 3 (2014): 825–840.

situation in which a foreign nation appears to be linked to several low-intensity intrusions centered on national and regional government information technology (IT) infrastructure. Fictional country designations were used, with a background describing a history of rivalry without major hostilities but including frequent cyber incidents between the countries. Information was presented in an intelligence-style packet containing media report excerpts and classified incident report snippets. The population was then presented with a high-level assessment of the rival's activities that included (1) a technical assessment of attribution confidence, (2) a possible strategic contextualization of actions observed, and (3) a concluding evaluation that events presented were (with medium confidence) likely precursors to a more serious cyber attack. All detail was held constant across participants, who were then asked to answer three questions. First, based on a preset list of policy responses, respondents offered a rank-order prioritization based on the options they felt would stand the greatest chance of deterring a more serious follow-on cyber attack by the rival state. Second, they selected (on a scale) how confident they were that their highest-ranked selection would successfully deter such an attack. Finally, they were asked to report from a preset list of options (with a write-in slot available) the primary reason they might request further information before proceeding.

The treatment in this experiment was in the presentation of intelligence. The population was randomly assigned to one of three groups in which intelligence analyses presented in the vignette took one of three forms: (1) entirely the result of human analysts; (2) coming from a defense department AI system with strategic contextualization and a final evaluation coming from human inputs; and (3) entirely generated by an AI system.

Study #2 consists of a similar study insofar as the population of respondents was split into three treatment groups based on the type of AI involvement in the scenario. However, the details differed to accommodate the alternative focus on the rival's use of AI for cyber activities. In the first of two phases, respondents were presented with a series of events similar to those in Study #1. The scenario broadly suggests a potential preparatory campaign laying the groundwork for a potential cyber-enabled disruption and political interference operation. Since the focus here is on AI in the loop, the goal with this design was simply to introduce a cyber conflict situation with a set amount of uncertainty and hold that constant across respondents.

After receiving scenario information, respondents are simply asked two questions: (1) whether they would like further information at that time, and (2) if yes, what kind of information they would like to see prioritized (from a preset list, including technical, strategic, and historical background). In a second phase, respondents are given additional technical and strategic information on elevated rival activities pointing

towards the same potential interference campaign. They are then presented with a list of preset policy options similar to those in Study #1. This list includes the option for additional intelligence requested, presented as the default option, as well as an array of responses ranging from private attribution to preemptive cyber operations and military mobilization. The presentation of a default option is significant, as it makes alternative choices an escalatory action. Respondents then select their preferred course of action and rate their confidence as to its appropriateness. An open field textbox is available here for them to offer a rationale if desired. Respondents are then permitted to select a second top preference alongside another confidence rating.

With Study #2, the treatment in this experiment was in the presentation of rival state cyber activities. For the first group (assignment was random), there are no AI inputs to the cyber conflict episode. Vignette descriptions clearly assign both tactical and strategic agency for the rival state's actions to hackers, defense planners, and political leaders. The second group was presented with a picture of rival cyber activities as increasingly enabled and determined by AI systems that chose techniques and specific targets under the auspices of well-established strategic objectives set by human agents (commanders, leaders, etc.). Finally, the third group was presented with an image of these activities conducted as the result of AI systems being used by the rival state to dictate strategic preferences. Here, information about tactical and operational agency is kept vague so as to avoid priming respondents towards assuming that AI is either involved or entirely omitted.

B. Findings for Study #1: AI in the Intelligence Loop

Both studies provide a unique and as-yet-unparalleled look into the role of AI in complex cyber conflict events. Several elements of the design of these studies are particularly worthy of note. Here I present the findings of Study #1. While more extensive testing is possible with this data, for space reasons, I initially present only the primary findings emerging from the core prompts respondents addressed following their exposure to the vignette. Table I below outlines respondent preferences for policy options presented to them following their exposure to the scenario vignette of a rival state appearing to lay the groundwork for a more significant disruptive attack. The table delineates responses across the control and two treatment groups. Specifically, Table I takes the top policy option choice for each respondent as the singular preference for an action response. Though the question was framed as a rank order preference exercise, the primary value in doing so was to avoid a situation where respondents felt that they had to select less assertive options because a single choice was being forced upon them (as opposed to a slate of potential options). Significantly, the two most favored options were private attribution to the rival state and a preemptive cyber operation aimed at setting a red line. Both Group #1 and Group #2 favored the latter option, a clear cyber strike intended to emphasize their state's objection to the groundswell in

activity. Group #3, by contrast, preferred private attribution or even the provision of economic sanctions as either superior or equal options to an in-domain cyber response.

TABLE I: RESPONSE SELECTIONS BY GROUP

	Wait & See	Private Attribution	Diplomatic Outreach	Public Attribution	Criminal Indictments	Economic Sanctions	Offensive Cyber Operation	Military Mobilization
Group #1 (Human Analysis)	9.1%	19.9%	9.3%	4.2%	7.4%	11.4%	31.2%	7.5%
Group #2 (AI Assistance)	8.5%	20.2%	7.1%	1.9%	12.6%	10.9%	34.3%	4.5%
Group #3 (AI Generated)	11.2%	27.4%	5.2%	0.9%	7.3%	24.2%	17.2%	6.6%

Table II presents confidence scores for the choices made above. Here, the point of greatest interest is the dramatically lower score average (on a 10-point scale, with 10 being the highest) for Group #3, at nearly 3 points lower than that of Groups #1 and #2.

TABLE II: CONFIDENCE SCORE BY GROUP

	Group #1 (Human Analysis)	Group #2 (AI Assistance)	Group #3 (AI Generated)
Confidence (Average)	7.2	6.92	3.98

An initial explanation for this might be found, of course, in the results from the final survey question in which respondents were asked to describe why they would be most likely to ask for additional information. As Table III shows, Group #1 and Group #2 show relatively similar scores for the inherent uncertainty of operations in cyberspace and the need for deeper geostrategic context as the driver behind any request for additional intelligence. Group #3, by contrast, cites the value of the intelligence assessment being presented as the principal source of uncertainty and need for further information.

TABLE III: INFORMATION REQUEST RATIONALE BY GROUP

	More Info Needed: Cyber Conflict Context	Intelligence Not Yet Credible	No Justification	Concerns for Assets	More Info Needed: Geostrategic Context	No Appropriate Options
Group #1 (Human Analysis)	16.9%	7.2%	17.5%	17.0%	32.2%	9.2%
Group #2 (AI Assistance)	19.3%	8.3%	15.5%	18.4%	25.6%	12.9%
Group #3 (AI Generated)	13.2%	26.5%	12.5%	14.1%	16.5%	17.2%

Table IV incorporates respondent information to attempt to better delineate the machinations for those faced with the vignette scenario. Specifically, it returns to the question of confidence in selected outcomes across the treatment groups and places the results in three categories: (1) length of time spent in a national-security career (education or profession), (2) cyber-specific know-how, and (3) foreign policy know-how.

TABLE IV: CONFIDENCE BY DEMOGRAPHIC DATA

		Confidence (Avg.)
Group #1 (Human Analysis)	Career Length	7.5
	Cyber Know-How	7
	Foreign Policy Know-How	6.8
Group #2 (AI Assistance)	Career Length	6.7
	Cyber Know-How	8.1
	Foreign Policy Know-How	6.2
Group #3 (AI Generated)	Career Length	4.3
	Cyber Know-How	3.7
	Foreign Policy Know-How	3.6

Trends among the population show that the more years spent in a national security career, the more certain those individuals in Group #1 are regarding their selected course of action. This holds, though less strongly, for Group #2. With the third group, however, there is yet again a distinct difference in stated perspective, with career experience appearing to matter little in predicting response confidence. Know-how, measured as index variables constructed from a set of pre-survey questions focused on assessing foundations of cyber conflict and foreign-policy-making knowledge,

likewise shows the same tendency for respondents in Group #3 to show less confidence regardless of deeper expertise. Of particular interest here, of course, is that know-how does otherwise appear to have some effect on confidence levels among respondents, as cyber-specific expertise is more positively associated with confidence in selected outcomes for Group #2—where AI systems are responsible for providing technical intelligence—than Group #1. As discussed below in the concluding section, it is the use of AI processes as a holistic tool that appears to sink respondent confidence in presented intelligence.

C. Findings for Study #2: AI in the Adversary Toolkit

Turning to the role of AI in adversary cyber activities and strategy, Tables V and VI show response results across the three treatment groups pertaining to Study #2's first phase. As Table V shows, a higher proportion of respondents in both groups exposed to AI elements asked for information than their Group #1 counterparts. Table VI then shows that interest in different kinds of information diverged across all groups. Group #1 was principally interested in additional background information on relations between the states in question, while Group #2, though also interested in further situational background, cited a need for additional event information and analysis. In contrast with both, Group #3—whose scenario outlined an adversary whose AI adoption centered on strategic coordination of state mechanisms of power and influence—overwhelmingly cited the need for more nuanced information on how emergent technology was being utilized by the rival state in addressing the present situation. Here, perhaps the most interesting comparison is that between Groups #2 and #3, where the tactical deployment of AI led to only a minimally elevated statement of need for further information on emergent technology usage over the baseline, but the presence of AI systems as strategic process produced a much greater effect.

TABLE V: INFORMATION REQUESTS BY GROUPS

	Further Information	
	Yes	No
Group #1 (No AI Inputs)	63.4%	36.6%
Group #2 (Tactical AD)	82.2%	17.8%
Group #3 (Strategic AI)	84.9%	15.1%

TABLE VI: INFORMATION REQUEST RATIONALE BY GROUPS

	Technical Detail (Event)	Political Situation (Event)	Technology Usage	Geopolitical History
Group #1 (No AI Inputs)	16.2%	32.3%	18.9%	32.6%
Group #2 (Tactical AI)	13.9%	35.4%	17.2%	33.5%
Group #3 (Strategic AI)	14.1%	23.3%	44.3%	18.3%

Study #2 at this juncture provides additional technical and event information to respondents, though the gist of the situation (i.e., an implied buildup towards a potentially more disruptive cyber campaign) remains static. Table VII below outlines respondent top preferences for policy options presented to them following their exposure in the second phase. There is a minor preference across all groups for additional intelligence, but notably so for those in Group #3, where AI plays a more strategic role. Otherwise, Group #2 respondents appear to be more assertive than their counterparts, opting for potentially escalatory policy actions like public attribution, economic sanctions, and counteroffensive cyber operations more consistently than those in either other group. Interestingly, Group #3 appears to feature the least assertive respondents, who preferred to privately attribute or wait for additional intelligence over alternatives. Across groups (Table VIII), confidence in the appropriateness of the selected response option is reasonably high, but particularly for Group #1.

TABLE VII: PRIMARY RESPONSES ACROSS GROUPS

	Additional Intelligence	Wait & See	Private Attribution	Diplomatic Outreach	Public Attribution	Criminal Indictments	Economic Sanctions	Offensive Cyber Operation	Military Mobilization
Group #1 (No AI Inputs)	7.9%	4.2%	18.2%	10.7%	20.4%	7.2%	11.1%	18.2%	2.1%
Group #2 (Tactical AI)	9.2%	1.9%	11.2%	3.9%	17.4%	11.3%	19.3%	21.5%	4.3%
Group #3 (Strategic AI)	16.3%	5.8%	23.6%	8.3%	14.4%	4.2%	9.2%	16.4%	1.8%

TABLE VIII: GROUPWISE CONFIDENCE SCORES (PRIMARY RESPONSES)

	Group #1 (No AI Inputs)	Group #2 (Tactical AI)	Group #3 (Strategic AI)
Confidence	8.3	6.6	6.8

Tables IX and X delineate secondary respondent preferences. The main deviation worthy of note here is the clear preference across all groups to default to further intelligence gathering as a fallback position in the crisis. This may imply that certainty in the ability to act is time-limited, as other studies have suggested might be true of cyber conflict decision-making, or that repeated low-intensity experiences with adversary behavior create an expectation of limited escalation potential.

TABLE IX: SECONDARY RESPONSES ACROSS GROUPS

	Additional Intelligence	Wait & See	Private Attribution	Diplomatic Outreach	Public Attribution	Criminal Indictments	Economic Sanctions	Offensive Cyber Operation	Military Mobilization
Group #1 (No AI Inputs)	19.4%	7.2%	18.2%	14.3%	15.2%	3.2%	7.9%	13.2%	1.4%
Group #2 (Tactical AI)	23.9%	7.7%	20.0%	7.1%	14.2%	3.4%	11.5%	11.1%	1.1%
Group #3 (Strategic AI)	27.0%	11.2%	23.6%	8.1%	7.2%	1.8%	13.2%	7.3%	0.6%

TABLE X: GROUPWISE CONFIDENCE SCORES (SECONDARY RESPONSES)

	Group #1 (No AI Inputs)	Group #2 (Tactical AI)	Group #3 (Strategic AI)
Confidence	6.1	5.5	5.2

As with Study #1, Table XI below presents confidence results for both primary and secondary choices across the three groups arrayed by demographic criteria. Here, length of time in a national security career is again associated with heightened confidence among respondents. Unlike Study #1, however, this effect broadly appears to hold across all three groups. At the same time, while a cyber-specific background appears to make respondents more generally confident of the appropriateness of their selected policy option, there appears to be no significant variation based on background between the groups.

TABLE XI: CONFIDENCE BY DEMOGRAPHIC DATA

		Confidence (Avg.)
Group #1 (No A1 Inputs)	Career Length	7.1
	Cyber Know-How	5.4
	Foreign Policy Know-How	5.8
Group #2 (Tactical AI)	Career Length	7.3
	Cyber Know-How	4.9
	Foreign Policy Know-How	5.2
Group #3 (Strategic AI)	Career Length	6.8
	Cyber Know-How	5.2
	Foreign Policy Know-How	4.8

Finally, the design of Study #2 provides a unique opportunity to assess the willingness to escalate or maintain assertive behavior across the two phases of the scenario. Table XII presents a logit regression analysis of utilizing both treatment and demographic detail about respondents. The first dependent variable is a simple measure of willingness to escalate from the first to second phase (i.e., escalation up the policy option ladder did or did not occur from either more intelligence gathering or doing nothing). This measure considers any policy option selected for phase two other than the same request for further intelligence to be escalatory in nature. The second policy option draws a line around the first four policy options (including doing nothing, private attribution, diplomatic engagement, and public attribution) as non-escalatory and then produces the same dichotomous account of escalatory willingness.

TABLE XII: BINOMIAL LOGIT ANALYSIS

	DV #1		DV #2	
	(Model 1)	(Model 2)	(Model 3)	(Model 4)
Human Only	**0.434 (0.282)	**0.372 (0.323)	**0.67 (0.299)	**0.55 (0.358)
Tactical AI	***0.529 (0.013)	***0.439 (0.029)	***0 89 (0.018)	***0.821 (0.033)
Strategic AI	0.212 _	0.19 (0.488)	0.321 (0.398)	0.239 (0.498)
Education	_ _	0.032 (0.334)	_ _	0.059 (0.401)
Career Length	_ _	**-0.589 (0.089)	_ _	**-0.694 (0.101)
Cyber Know-How	_ _	**-0.321 (0.069)	_ _	**-0.252 (0.078)
FP Know-How	_ _	*0.112 (0.201)	_ _	*0.143 (0.280)
Risk Aversion	_ _	*0.456 (0.157)	_ _	*0.958 (0.171)
Observations	503	503	503	503
L1	-242.118	-238.189	-245.238	-229.832
Pseudo R^2	0.089	0.101	0.092	0.104

Note: * p<0.05; ** p<.01; *** p<.001

The choice to escalate actions between the phases was fairly common except among the 503 respondents for Study #2, as was the selection of more assertive policy options during the second phase. As Table XII shows, willingness to escalate (particularly using the second dependent variable (DV) measure in Model 2) is most strongly predicted by being in both Group #2 and, to a slightly lesser extent, the control group. This suggests that the presence of AI in shaping rival cyber activities introduces uncertainty sufficient to produce restraint and hesitancy among decision-makers but that the use of AI for technical advantage is seen as a provocation worthy of a response. Likewise, career length and cyber know-how are both negative predictors of escalatory preferences, while foreign policy know-how is slightly positively predictive thereof (but only with weak significance). Taken together, these findings present a unique picture of how AI in the decision-making loop interacts with decision-maker's predispositions and the broader contours of cyber conflict to affect views on the value of information and the appropriateness of using force.

4. DISCUSSION

The results of these studies on the inclusion of AI in the decision-making process of cyber conflict episodes provide a unique set of insights for those invested in building effective technical, institutional, and cultural programs for national-security AI adoption.

A. Evolution vs. Revolution across Levels of Analysis

A notable takeaway from these studies is that AI augmentation of the decision-making process at the level of technical capabilities/analysis appears to have only a limited impact on the inclinations of decision-makers. Results show a higher willingness to escalate perceived hostilities and even punish diffuse cyber aggression where AI is being used tactically. At first glance, this suggests that respondents see AI as an evolution of present conditions rather than a transformation of battlespace fundamentals.

By contrast, decision-makers seem to show restraint when AI is applied more holistically. Significantly, this applies to the picture being presented to decision-makers and the picture drawn for decision-makers by adversary actions. In debriefing, respondents clearly considered any holistic use of AI for intelligence (i.e., to include cross-domain assessments, ascription of a rival state's intentions, and highly-specific action recommendations) extremely premature. More than 24% of all respondents across both studies made reference to the potential error proneness of AI systems. This clearly leads many to question the macro-foundations of the patterns they see when the assumption that human agency principally drives strategic deployments disappears.

B. Robustness vs. Accuracy Assumptions

Across both studies, a sizable minority of respondents assumed that both foreign and homespun AI systems would be designed to be robust rather than perfectly accurate. The robustness vs. accuracy tradeoff is a well-known challenge in designing machine learning systems.[37] Robust systems are those that are hard to deceive, preferring to occasionally misclassify an observed variable so as to most effectively guarantee that no failure to capture legitimate inference is missed. By contrast, accurate systems are those that more often correctly classify data with limited misclassification, often at the expense of failing to classify in the face of attempted deception or outlier cases. The tradeoff in designing AI systems is that as one is made more robust, it tends to become less accurate and vice versa.[38] In debriefing, a range of respondents reasoned that any system trusted with strategic analysis would have to exhibit a certain error

37 Wyatt Hoffman, "Making AI Work for Cyber Defense" (CSET, 2021), https://doi.org/10.51593/2021CA007.

38 Ibid.

proneness. The alternative would be missed comprehension of the cues, signals, and developments that characterize the landscape of international affairs. Several respondents found this reasonable but pointed out that this uncertainty might be more novel than the usual human error assumed to be a part of foreign policy machinations, thus inviting greater-than-usual restraint on their parts.

C. The Perception of Human Control in the Loop

While there is clear evidence that respondents felt that holistic AI usage might be premature and error-prone, results also seem to point to a willingness to accept AI-produced information so long as human agency was presented as guiding the process. While decision-maker confidence is clearly affected by AI inclusion, findings show that this may be partially resolved by the perception of humans in the loop simply benefiting from some automated tool. Indeed, various respondents in Study #1's Group #2 stated that this kind of automated intelligence generation was not a far cry from basic data-collation tools they were already used to.

Taken at face value, this finding is concerning for those invested in building effective AI systems to bolster national security performance. If human agency in the loop creates a verification effect, then there clearly exists the possibility that some decision-makers will experience a false sense of validation of their assumptions without appropriate thought given to the credibility of outputs being presented to them. This possibility is particularly worrisome given that these results suggest that a clear view of humans in the loop (vis-à-vis some level of machine agency) is tied to greater assertiveness in the actions selected by respondents.

D. Prior Knowledge, Expertise, and AI Responsibility

Interestingly, pre-existing know-how seems to do little to bolster confidence in the choices made by respondents vis-à-vis holistic AI presence. True, technical background and experience (measured by the length of time individuals have been in their national security career) both seem to be linked to restraint. But there is relative consistency across results in Study #2 for respondents of all skill levels and backgrounds to prefer further information assessment when presented with AI-enabled adversary activities.

Here, open-ended responses and debrief sessions suggest that respondents of all stripes are not necessarily concerned that they are witnessing a sea change in approach to cyber engagement. Rather, respondents tended to express anxiety about AI systems—both homegrown and adversary—getting something wrong. This was expressed beyond the more nuanced criticism several respondents expressed about the potential proneness to error of even well-designed systems. Rather, numerous (n=38) respondents argued that human error was likely behind some AI shortcoming. Of these respondents, three specifically addressed accountability, stating that such errors

were not the responsibility of states to provide for as an allowance in international engagements but that errors inevitably invited additional strategic ambiguity.

5. IMPLICATIONS AND CONCLUDING REMARKS

The impact of AI systems and machine learning on the shape, scope, and aims of sophisticated cyber operations is a question of great interest among scholars and practitioners focused on issues of global cyber conflict. This initial report of two studies, however, provides some promising foundations upon which further testing and programming might be constructed.

Perhaps foremost among the implications here is the clear need to institute broad-scoped training on the fundamentals of AI precepts and functionality. For both studies, I attempted to gauge AI know-how among respondents alongside domain-specific and foreign-policy knowledge. However, so few participants scored above a low grade that the variable was (after initial testing) rejected for its insignificance. Clearly, domain-specific cyber knowledge had some impact on respondent objectivity, so it only stands to reason that knowledge about how AI works might limit the chances of bias impacting the decision-making process. How information about AI is presented is also significant in shaping elite preferences for more or less assertive response options. And it is particularly concerning that AI assessments or actions were considered to be similar to human-produced equivalents as long as the holistic picture was seen to come from human agency. There are a number of potential pitfalls to be found in such an effect if appropriate training, language, and process cannot be implemented to build bulwarks against unwanted bias.

For IR theorists and international security researchers more broadly, these studies clearly link emergent technologies to enhanced uncertainties in foreign policy engagements. Here, these findings suggest that the upgrading of cyber conflict activities via AI may not elicit sufficient concern from operators and decision-makers who see AI tactically employed within known human-derived frameworks. Operationally, the idea that flaws in how AI systems work are more the responsibility of competitors than a point of strategic consideration for decision-makers is also quite concerning, as it implies a specific mechanism by which signaling in cyberspace is to become more difficult. This is especially worrisome given some countries' doctrinal orientation towards cyber deterrence by offensive action. And, of course, this study reinforces the idea that an expansion of cyber conflict processes for the purposes of better understanding and potentially poisoning AI systems is likely. After all, security institutions seeking to clarify the role played by AI in shaping observed contestation will inevitably work to produce a more complete picture of how algorithms and human

agency are being set up to improve performance. Given these implications, continued work on building foundations for a better understanding of the prospective impact of AI on the future contours of global cyber conflict seems vital.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

Subverting Skynet: The Strategic Promise of Lethal Autonomous Weapons and the Perils of Exploitation

Lennart Maschmeyer
Senior Researcher
Center for Security Studies
ETH Zurich
Zurich, Switzerland
lmaschmeyer@ethz.ch

Abstract: Lethal Autonomous Weapons Systems (LAWS) promise a revolution in warfare by increasing the lethality of force while reducing the costs of war. Yet these gains come at the cost of significant yet underappreciated perils. LAWS are vulnerable to subversion, allowing adversaries to degrade or disable these systems or even turn them against their makers. Subversion involves exploiting flaws in complex systems to make them behave in unexpected ways. It is possible because, like other computer systems, the behavior of LAWS is determined by logical rules and routines. These rules and routines inevitably contain flaws, creating vulnerabilities that adversaries can exploit. This paper identifies three avenues of subversion: (1) manipulating the algorithm itself during the design process, (2) poisoning the data used to train the artificial intelligence operating the LAWS, and (3) manipulating physical objects LAWS are trained to respond to. This potential for subversion creates fundamental uncertainty for strategic planners, military commanders, and soldiers in the field. LAWS are powerful capabilities that may win wars, yet they may also become liabilities that lead to defeat against crafty adversaries. Hence, fulfilling the strategic promise of LAWS requires mitigating their vulnerabilities. Examining possible mitigations, the paper shows that technical fixes are currently unavailable, necessitating strategic solutions. It identifies two possible solutions. The first is employing counterintelligence strategies and tactics to detect, neutralize, and pre-empt attempts at subversion. The second is adopting a force structure model that maintains human superiority to neutralize rogue LAWS if necessary. However, both solutions reduce operational effectiveness and the strategic value of LAWS, thus

forfeiting some of the core advantages these new systems promise. Consequently, the paper concludes that the strategic challenges of deploying LAWS currently outweigh the opportunities, necessitating a cautious approach and a greater prioritization of strategy development.

Keywords: *autonomous weapons, artificial intelligence, strategy, cybersecurity, subversion, exploitation*

1. INTRODUCTION

There have been many revolutions in warfare, but few with as much transformative potential as autonomous weapons. Current expectations envision that Lethal Autonomous Weapons Systems (LAWS) will hold the key to success in future war, enabling unprecedented gains in military power (Horowitz 2018; Scharre 2018b; Gill 2020). LAWS are expected to provide vast improvements in the speed, efficiency, and effectiveness of military operations (Payne 2018; Horowitz 2019). Development of different forms of Artificial Intelligence (AI), primarily neural networks, is seeing exponential gains in their capacity to learn, process information, and outperform humans in complex tasks (AlphaStar Team 2019; Zhou et al. 2021). Consequently, many expect LAWS powered by such machine intelligence to process more information faster than humans while remaining immune to exhaustion and emotion (Umbrello, Torres, and De Bellis 2020; Brooks 2015).

Accordingly, current expectations see vast strategic promise. Russian president Vladimir Putin has gone as far as to suggest that "whoever becomes the leader in this sphere will become the ruler of the world" (Gigova 2017). Accordingly, NATO identifies the acquisition of AI and autonomous weapons as "fundamental to the future security of NATO and its Allies" (NATO 2020, 29). The military revolution these capabilities promise, moreover, is not merely some possibility for the distant future. In fact, some suggest it may be right around the corner. As former US Navy Secretary Ray Mabus stated in 2015, the ultramodern F-35 fighter jet "should be, and almost certainly will be, the last manned strike fighter aircraft" (Myers 2015).

Considering the speed of development and the expected strategic potential of LAWS, it is worth considering the pitfalls of deploying them in practice. Because LAWS capabilities have not yet been fully realized, analysis of this type inherently involves a speculative component. However, existing research in computer science and political science makes it possible to identify both the likely architecture of LAWS and the types of vulnerabilities that ensue, as well as the resulting strategic challenges. The starting

point for my argument is Stephen Biddle's key insight that technology itself does not determine political outcomes or victory in war (Biddle 2010). Existing research indicates that deploying LAWS involves challenges that are as great, if not greater, than their promise. There are three main challenges. First, despite the vast progress in AI, it has important shortcomings that will likely prevent LAWS from fulfilling the vast expectations, particularly concerning autonomous decision-making and judgment (Pietrucha 2015; Goldfarb and Lindsay 2022). Second, these shortcomings and the lack of transparency of decision-making processes limit reliability and hamper effective human-machine interaction (Horowitz 2019, 783; Scharre 2018a). Third, there is the looming ethical question whether "killer robots" are ever morally acceptable ("Stopping Killer Robots" 2020). In short, to reap the rewards LAWS promise, actors must overcome significant obstacles that require further technical development, careful ethical reasoning and doctrinal innovation. These challenges are steep but not insurmountable. However, even if these challenges can be overcome, I argue that there remains a more fundamental, underappreciated peril.

Specifically, I focus on a class of AI algorithms whose advantages in information processing and decision-making make them the most likely form of AI to be used in LAWS, namely neural networks (Scharre 2018b; Vazquez 2019; Zhang et al. 2021). Importantly, neural networks are inherently vulnerable to subversion, meaning adversarial exploitation of vulnerabilities in their design that allow hostile manipulation of their actions. Research in computer science identifies three main techniques of subverting neural networks: (1) development compromise, (2) data poisoning, and (3) input manipulation. I discuss each type of attack in detail and show how, in the worst case, subversion of this kind allows whole armies to be turned against their makers. However, even more subtle manipulations can neutralize the strategic advantage of LAWS, turning capabilities into liabilities. Existing strategic visions and analyses fail to adequately consider this danger. For example, NATO's 2030 vision stresses the need to "exploit the power of AI-driven technologies" (NATO 2020, 30) but not the risk of adversaries doing the same only to turn these technologies against their users. Moreover, even short of such catastrophic scenarios, the mere potential for exploitation undermines effectiveness. Once this potential is demonstrated—a question not of if but of when—how can soldiers trust LAWS to have their back? The result is lingering uncertainty over whether LAWS will in fact fight for one's side, which hampers effective deployment and undermines strategic value.

Importantly, these vulnerabilities are inherent to the design of AI techniques and thus cannot be fully eliminated (Comiter 2019). Hence, the solution must be strategic. The paper identifies two strategic solutions. The first is building on counterintelligence strategies to detect, neutralize, and pre-empt adversary subversion. The second is maintaining a force structure that keeps superior firepower in the hands of humans

to enable a "Luddite option" of taking out rogue LAWS if necessary. Both strategies, however, reduce the efficiency and effectiveness of LAWS deployment, which risks forfeiting their strategic promise. Consequently, I argue that the challenges involved in deploying LAWS currently outweigh the opportunities.

To build this argument, the paper proceeds in three steps. First, integrating insights from international relations scholarship with current research in AI, it explains why the design of LAWS makes them vulnerable to subversion. Second, based on demonstrated paths of subversion, the paper identifies the various threat types faced by LAWS. Third, it examines possible technical and strategic solutions to this problem. Finally, it examines the geopolitical implications of this situation.

2. SUBVERTING TECHNOLOGY: VULNERABILITIES, EXPLOITATION, AND UNCERTAINTY

Like all computer systems, LAWS are susceptible to subversion. LAWS in general refer to weapons systems capable of lethal effects that operate autonomously. The US Department of Defense defines them as "a weapon system that, once activated, can select and engage targets without further intervention by a human operator" (US Department of Defense 2017, 13–14). Such weapons do not (yet) exist, and as mentioned, some oppose their creation on moral grounds. Nonetheless, there are no legal prohibitions on them, and a recent report by the United States Congressional Research Service notes that "U.S. policy does not prohibit the development or employment of LAWS" (Sayler 2020). Considering their strategic promise, it is likely only a matter of when, not if, such capabilities are developed.

Specifically, this paper focuses on (still speculative) future weapons systems powered by advanced AI capable of fully autonomous "decision-making" and movement on the battlefield. In a popular definition, AI refers to "the science and engineering of making computers behave in ways that, until recently, we thought required human intelligence" (High 2017). AI itself is an umbrella term for different types of machine learning techniques, from relatively straightforward pattern-matching algorithms to complex "neural networks" (Kozma et al. 2019). Neural networks are complex algorithms that replicate the structure of the human brain as well as its learning process—but only to some extent (more on that below) (Bose and Liang 1996). Due to their ability to autonomously process information, learn and make "decisions," neural networks are the most likely type of technique to be used in LAWS. They offer great promise of emulating aspects of human intelligence while surpassing human capabilities in key regards, above all the speed and quantity of information processing (Payne 2018, 8). For example, neural networks have been getting better than human

experts at diagnosing diseases such as cancer based on image recognition (*Science Daily* 2021; Lee 2021). Meanwhile, Google DeepMind's "AlphaZero" neural network learned complex games by playing out vast numbers of iterations against itself and identifying winning strategies. In doing so, it beat human world champions in complex challenges such as the strategy game Starcraft II (AlphaStar Team 2019). Despite their seemingly "intelligent" properties, however, functionally LAWS are simply advanced computer systems controlling physical hardware through complex algorithms.

Computer systems have enabled vast efficiency gains in practically all fields of human endeavor. In particular, the information revolution in warfare has significantly improved the speed, lethality, and effectiveness of military units (Adamsky and Bjerga 2010; Lindsay 2020). However, computer systems also produce new liabilities because they involve vulnerabilities. Security scholars who focus on cybersecurity have examined the geopolitical role of the exploitation of such vulnerabilities to achieve political outcomes. Prevailing wisdom holds that "cyber operations" employing this mechanism are new instruments of power (Kello 2013; Fischerkeller and Harknett 2017; Buchanan 2020). More recent work, however, shows that their reliance on exploitation reveals cyber operations to be instruments of subversion, a classic mechanism of power in intelligence operations. Subversion involves the exploitation of vulnerabilities in systems of rules and practices to gain access and undermine or manipulate these systems to produce outcomes not intended by their designers, owners, or participants (Maschmeyer 2021, 54). Traditional subversion uses spies to infiltrate and exploit social systems, while cyber operations infiltrate and exploit computer systems (ibid.). In principle, because humans design these systems, and because humans are fallible, all computer systems are susceptible to hacking. Accordingly, vulnerabilities have been found even in advanced military hardware where security is a key priority—the F35 fighter jet is a key example (*Global Defence Technology* 2019). Importantly, as previous research has shown, AI techniques are vulnerable to exploitation as well (Comiter 2019; Whyte 2020).

In short, the inevitable presence of vulnerabilities in computer systems enables subversion. The resulting potential for subversion creates uncertainty. Since any system could plausibly be subject to adversarial manipulation, no system can be fully trusted. The more capable computer systems are of causing harmful effects and destruction, the greater the potential for unexpected losses. As advanced computer systems operating machinery with unprecedented lethality, LAWS produce unprecedented risks. Moreover, as the next section will show, the design characteristics of AI exacerbate their vulnerability.

3. PATHS OF SUBVERSION IN LAWS AND RESULTING THREATS

Importantly, the same characteristics that enable the advanced information-processing and decision-making capabilities of neural networks also provide vulnerabilities that allow adversaries to subvert them. Neural networks consist of complex algorithms capable of processing far more information and making "decisions" based on this analysis much faster than humans. Yet despite being roughly modeled on neurons in a brain, these algorithms function fundamentally differently; moreover, their functioning remains opaque and not fully traceable. Hence they are also known as "black boxes" (Bathaee 2017). Consequently, crafty adversaries can manipulate neural networks to behave in ways not intended by their designers and users, while the manipulation remains hidden and difficult for humans to detect. Neural networks are vulnerable to three main techniques of such subversion, known as "adversarial attacks" in computer science: (1) development compromise, (2) data poisoning, and (3) input manipulation.

Development compromises manipulate the AI algorithm itself, enabling malicious actors to degrade its performance, render it useless, or produce a range of unexpected outcomes. As Tyukin et al. (2021) show, such manipulation can require minimal changes—up to only a single artificial neuron in the network—in order to change trigger conditions that produce unexpected responses to a given input parameter. In the context of LAWS, imagine a robotic ground soldier suddenly refusing to follow orders after being exposed to a seemingly innocent sunflower pattern or an inaudible sound produced by the opposing army. This risk is not theoretical. Recent research demonstrates the practicability of inserting such "backdoors," even showing the feasibility of spreading them en masse by combining them with traditional computer viruses (Chen et al. 2017; Qi et al. 2021). Importantly, because the way neural networks produce output remains a black box, identifying and removing vulnerabilities by analyzing code is not possible, as it is with "traditional" software. This situation offers a key advantage for adversaries: manipulation remains undetectable before the trigger activates the malicious action(s). The neural network operates as expected, until, unexpectedly, it does not. A second advantage is the wide range of effects this direct manipulation enables.

However, direct manipulation also requires direct access to the development process. Tyukin et al. (2021) propose a "disgruntled software engineer" as a possible attack vector, underlining the linkage to classic intelligence operations seeking vulnerable employees at targeted organizations to "turn" them—that is, secretly convert them to work for one's own side (Andrew 2000, 232).

The second form of subversion, data poisoning, involves manipulating data used

while training the neural network. It allows shaping the model a neural network produces for information analysis, processing, and response. For instance, when training an image recognition neural network, a set of sample images with appropriate classifications is fed into the model as a baseline. Adversaries can manipulate either the classification or the image content itself. Manipulating the classification data is relatively easily detected by human supervisors and thus not very reliable. The same does not apply to manipulations of the image content. The latter type of manipulation is known as introducing "adversarial examples," which "enable adversaries to subvert the expected system behavior leading to undesired consequences and could pose a security risk when these systems are deployed in the real world" (Narodytska and Kasiviswanathan 2017). Importantly, the manipulation can remain entirely invisible to human supervisors.

As Sheu (2020) underlines, neural networks are "capable of making accurate predictions based on 'peripheral' features or noise but that contain no heuristic or scientific meaning other than statistical correlation with the labels." Crafty adversaries can exploit this characteristic by manipulating only a few pixels in images, invisible to the human eye yet producing statistical correlations influencing prediction (Narodytska and Kasiviswanathan 2017). Athalye et al. (2018) demonstrated how this exploitation made Google's powerful image recognition neural network identify a turtle as a rifle and vice versa. The implications for LAWS are obvious: imagine a synthetic army misidentifying an opposing army as holding turtles—or, conversely, wiping out a turtle colony because it looks like an arms cache.

Exploiting this vulnerability requires access to training data, however. Provided that data is collected and processed in-house, doing so would likely require a human in the loop who works at the organization involved. Accordingly, counterintelligence strategies could help thwart such subversion, or at least minimize their risk. However, in practice, the bigger the dataset used for AI, the better the model tends to become (Hestness et al. 2017), and some of the largest datasets are public (Towards AI 2021). Hence, foregoing available data and (re-)collecting relevant data will add to development cost. For powerful government agencies, this problem is surmountable. Nonetheless, the larger the scale of the data collection, the more possible points of compromise there are. Subversion remains a threat.

The third vulnerability, input manipulation, exists because of the opaqueness of analysis algorithms and their shortcomings, allowing adversaries to manipulate input data to produce unintended outcomes. As discussed, neural networks recognize inputs based on pattern models. With sufficient training, these pattern models can identify the objects they are trained for as well as or better than humans. However, the process of recognition works fundamentally differently than human cognition (Bezdek 1992).

Because of this difference, it is possible to manipulate objects in ways that confuse the neural network yet remain invisible or not readily apparent to human cognition. Research by Eykholt et al. (2018) demonstrates that minor physical changes to objects, such as adding stickers to road signs, can confuse pattern recognition models enough to cause total recognition failure. As McAfee researchers have demonstrated, such manipulation can be subtle enough to evade human eyes yet trick autonomous vehicles into misidentifying a 35 mph speed sign as an 85 mph sign (Keck 2020).

Once flaws in the pattern recognition algorithms are identified, this vulnerability requires less technical sophistication to exploit than the previous two vulnerabilities discussed above. Nonetheless, it enables similarly effective ways to subvert LAWS. For example, imagine an adversary army wearing uniforms containing visual patterns that make its units appear, to the LAWS, to be non-combatants or friendly forces. Moreover, the patterns need not be visual. Consider a hostile force transmitting an audio signal inaudible to humans, yet detected by the LAWS, which identify it as the sound of an incoming sandstorm and shut down and take a protective stance in response. In an extreme case, such signals could contain commands to the LAWS that allow their takeover. This mechanism has already been demonstrated in practice with digital assistants (Lei et al. 2018).

In short, existing research indicates that future LAWS will almost certainly involve significant, potentially catastrophic vulnerabilities. Exploiting them in practice, and without alerting the victims to the fact that such exploitation has taken place, is undoubtedly challenging. Even cyber operations against existing, non-intelligence computer systems face important constraints that limit effectiveness (Maschmeyer 2021). However, compared to the challenges of exploiting "non-intelligent" traditional computer systems, at least two of the exploitation techniques for neural networks just discussed have lower demands. While hacking requires establishing access to complex systems themselves, neither data poisoning nor input manipulation do. Accordingly, Nguyen et al. (2015) show it is "easy to produce images that are completely unrecognizable to humans, but that state-of-the-art DNNs [deep neural networks] believe to be recognizable objects with 99.99% confidence." Moreover, as discussed above, given sufficient knowledge about the target system, they are relatively simple to carry out—at least compared to, for example, complex hacks of industrial control systems (Lindsay 2013). In short, subverting LAWS is a challenge, but not a greater challenge than those that traditional cyber operations face. Meanwhile, the potential payoff is significant. If LAWS become a reality, actors could be expected to focus significant resources on finding ways to subvert them. This incentive, combined with the presence of vulnerabilities, produces fundamental uncertainty in their reliability.

4. STRATEGIC IMPLICATIONS

The possibility of subverting LAWS outlined above causes significant strategic challenges. Apart from the known mechanisms discussed above, the history of cybersecurity indicates that as adversaries experiment, new types of vulnerabilities will emerge. The result is fundamental uncertainty that creates a strategic dilemma. LAWS promise great gains in speed, lethality, efficiency, and effectiveness. Hence, governments have great incentives to adopt and deploy them. Militaries can also be expected to thoroughly test LAWS to reduce their vulnerability towards subversion. Yet as the case of the F-35 jet illustrates, even far less complex computer systems in existing military hardware contain vulnerabilities despite security being a key priority during development. Moreover, subversion is by definition unexpected, as it creatively uses existing logical rules and routines to produce expected outcomes. Consequently, even with thoroughly tested equipment, commanders can never be fully certain LAWS will operate as expected. Of course, in principle the same applies to armies consisting of human soldiers. However, the fundamental difference is the potential for systemic subversion. If a vulnerability exists in LAWS, it affects not only individual units but entire classes of systems. In the worst case, the result is entire synthetic armies shutting down or switching sides.

Considering this uncertainty, strategic planners might not want to count on LAWS for survival. The rational and prudent response would be to limit the number of fully autonomous LAWS and use them to pursue lower priority, non-vital strategic objectives. However, this approach risks giving adversaries fielding overwhelming numbers of LAWS a potentially insurmountable advantage. Against such an adversary, however, subversion creates the potential for victory without a battle if one identifies a suitable vulnerability in its LAWS. Using subversion involves another pitfall, however, since it is itself fraught with uncertainty. Because subversion involves the manipulation of highly complex and incompletely familiar systems, there is always a non-zero chance it will fail to produce the intended effect, or produce unintended consequences (Maschmeyer 2021).

To face these challenges, actors must prioritize two objectives: first, reducing uncertainty; and second, developing effective strategies under uncertainty. To reduce uncertainty, in theory the best solution is a technical fix to prevent exploitation. In practice, however, the vulnerabilities that allow subversion are inherent in the characteristics of currently conceivable forms of AI. Consequently, as Marcus Comiter (2019, 1) explains, "unlike traditional cyberattacks that are caused by 'bugs' or human mistakes in code, AI attacks are enabled by inherent limitations in the underlying AI algorithms that currently cannot be fixed."

A speculative solution is a developmental leap towards something closer to human intelligence. The more similar AI becomes to human intelligence, the more apparent attempts at exploitation will become to human handlers. However, such AI remains hypothetical. Moreover, it opens an entirely new can of worms. The obvious peril is the rise of self-improving superintelligence that rapidly surpasses any human capabilities, holding our species' survival in its artificial hands (Bostrom 2014). The movie *The Terminator* (Cameron 1984) and its sequels vividly illustrate one possible worst-case scenario involving a self-aware, nuclear-armed superintelligence called "Skynet." This route towards fixing the potential of subversion at the technical level thus poses its own dilemma: making existing models immune to subversion is impossible, while driving development towards human-like intelligence capable of resisting it makes Skynet scenarios possible.

With technical fixes out of reach for the foreseeable future, the next best available solution is strategic. Specifically, counterintelligence strategies can help improve the detection of adversary subversion and neutralize it by removing relevant vulnerabilities. As mentioned above, humans remain key in development compromises. Hence, employees at organizations tasked with development are possible targets for exploitation. The espionage literature is ripe with mechanisms of exploitation, such as using existing grievances, leveraging loneliness, or applying blackmail (Blackstock 1978; Herman 1996; Andrew and Mitrokhin 1999). Countermeasures include stronger vetting, security practices designed to thwart attempts at infiltration, and tracking of job satisfaction—apart from making sure employees are satisfied in the first place—as well as monitoring of suspect employees. Leaning into a counterintelligence mindset brings its own perils, however. Apart from consuming additional resources, it undermines institutional cohesion by producing "the special counterintelligence mentality: slightly paranoid, considering the possibility of manipulation and deception everywhere" (Herman 1996, 197–198). Hence, it risks undermining the effectiveness of the design process.

There are possible technical aids to detect compromises at the development stage as well. For example, recent research develops techniques to detect development compromises in neural networks (Bilgin 2021). Such research is in its infancy, however, and there are no current indications it will reliably fix the problem. Meanwhile, established counterintelligence strategies fall short of providing proven countermeasures for the novel types of vulnerabilities inherent in AI. The opaque nature of neural networks hampers the ability to anticipate possible exploitation via poisoned data or (physical) manipulation of input data. By their very nature, exploits remain unexpected until they are used. One can aim to eliminate possible ways to exploit vulnerabilities in one's LAWS through "read teaming" exercises, as in existing

cybersecurity solutions. Yet there are no guarantees these exercises will identify the same vulnerabilities as creative adversaries. Uncertainty remains.

The third solution addresses this uncertainty head on with a mixed force structure, maintaining superior firepower within human hands. In doing so, this strategy maintains a "Luddite option" of humans taking down marauding machines. Emerging strategies for the use of LAWS, such as the US Department of Defense's "Third Offset," already envision integrated force structures pairing humans and machines in what has become known as the "centaur model" (Freedberg 2015). Considering the uncertainty posed by the vulnerability of LAWS towards subversion, a prudent strategy would be to ensure that human troops always possess superior firepower. In the worst case of a hostile takeover of LAWS, as well as isolated cases of rogue synthetics marauding across the battlefield, human troops would thus be able to neutralize the threat. This strategy has several disadvantages, however. First, even if the human troops prevail in the worst-case scenario, there will still be human losses. Second, the need to maintain human superiority limits the deployment of LAWS at scale, thus offsetting one potential key advantage of machine armies. Third, it also limits the lethality of LAWS, which offsets their potential advantage of fielding highly mobile units carrying weapons too heavy for human soldiers. This strategy is thus suboptimal at best but nonetheless provides the best available solution for the problem of uncertainty.

There is a more elegant technical solution using the same mechanism as the Luddite option, i.e., humans neutralizing errant LAWS. Rather than using weapons, it would involve equipping LAWS with physical kill switches that instantly disable them and are easily accessible by nearby humans. In theory, if a LAWS behaves erratically, soldiers could just push a button and shut it down. In practice, however, there is a high chance of losses resulting in the worst-case scenario of LAWS turning hostile, since the soldiers would need to come close enough to touch these lethal synthetic units. Moreover, a key advantage of LAWS is their superior speed and mobility, thus potentially precluding humans from easily accessing the kill switch—particularly for airborne LAWS. As an engineer might point out, there is an alternative—a wirelessly activated kill switch. However, as with most technical solutions that increase convenience, there is a corresponding decrease in security. Namely, crafty adversaries who work out the frequency, command sequence, or trigger for the wireless shutdown would suddenly have a tool to disable armies at the push of a button. And surely, if there is such a kill switch, obtaining that information will become an adversary's number-one intelligence priority. In short, the more elegant technical solution involves drawbacks similar to those of the Luddite option.

Moreover, even without adversary subversion of LAWS in action, the mere possibility of such subversion is likely to undermine the operational effectiveness of mixed centaur units. Roff and Danks (2018) persuasively show why and how the unique features of LAWS, primarily the opaqueness of their cognition and decision-making mechanisms, preclude the formation of trust bonds between human and machine units, even if the LAWS function flawlessly. Now add the potential for subversion into this calculation. If soldiers already distrust their machine "partners" when they function well, how will these soldiers behave knowing there is a non-zero possibility that adversaries will manipulate or take over these partners? Since LAWS may behave in unpredictable ways even when working as designed, incidents of soldiers taking out LAWS out of simple fear of manipulation are inevitable. Moreover, there is a real risk of soldiers doing so at scale if panic spreads.

To conclude, there are currently no simple solutions to the problem of uncertainty. Technical fixes remain elusive, counterintelligence strategies have their limits and introduce new perils, and the extreme "Luddite option" of taking down rogue LAWS by force risks both significant losses as well as forfeiting the key advantages of fielding LAWS in the first place.

5. CONCLUSION

This paper has argued that LAWS are vulnerable to subversion, allowing adversaries to degrade, disable, or manipulate these systems. In the worst case, crafty adversaries could even turn LAWS against their makers. Subversion is possible because, like other computer systems, the behavior of LAWS and the AI algorithms that would be operating them is determined by logical rules and routines. These rules and routines inevitably contain flaws, creating vulnerabilities that adversaries can exploit to make the system behave in unexpected ways. Unlike existing "non-intelligent" computer systems, these flaws in logic cannot be easily detected and patched. Rather, the vulnerabilities that enable subversion of LAWS are inherent in the nature of AI and difficult to detect because AI algorithms tend to be "black boxes." There are three avenues of subversion: (1) manipulating the algorithm itself during the design process, (2) poisoning the data used to train the AI model, and (3) manipulating the physical objects the AI-enabled system is trained to respond to.

The potential for subversion creates fundamental uncertainty for strategic planners, commanders, and soldiers in the field relying on LAWS to win wars. The paper then examined possible mitigations for this uncertainty. It showed that technical fixes are unavailable, necessitating strategic solutions. There are two key solutions. The first is employing counterintelligence strategies and tactics to detect, neutralize,

and pre-empt attempts at subversion. The second is a centaur force structure model that maintains human superiority to neutralize rogue LAWS if contingencies arise. Both strategies are suboptimal because they reduce operational effectiveness and the strategic value LAWS can deliver—in the latter case, forfeiting the core advantages these new systems promise.

This argument leads to several important implications for doctrine, policy, and scholarship concerning the use of LAWS in battle. First, it is crucial not to overestimate their capabilities and predict strategic potential based solely on technological properties and possibilities. As this analysis has shown, not only do LAWS generate significant opportunities to enhance military power and gain advantages over adversaries, but deploying them also produces critical and possibly fatal challenges.

This situation leads to the second implication, namely that strategic challenges currently outweigh strategic opportunities because of the former's potential gravity. Despite the great promise of LAWS operating with unprecedented speed, efficiency, and lethality, their susceptibility to subversion turns this capability into an unprecedented liability in war. Strategic planners and policy-makers willing to open this Pandora's box should, at the very least, seriously engage with this danger and the non-trivial challenges involved in mitigating its impact.

Finally, without reliable solutions to the problem of subversion, the most prudent approach to the development and deployment of LAWS is caution. As awesome as their promise may be, the perils involved are dire. Developing LAWS with superior speed, lethality, and efficiency than human-controlled capabilities and deploying them at scale can certainly crush unprepared adversaries. Against prepared adversaries, however, these LAWS might become agents of one's own defeat. To avoid this fate, it is crucial to separate fiction from facts and speculation from analysis. It is vital to study the weaknesses of LAWS as much as, if not more than, their strengths and prioritize strategy development on an equal footing with technological innovation. LAWS may well enable a revolution in military affairs. Yet without caution and attention to strategy development similar to that given to technological development, the outcome of that revolution might be the opposite of what its advocates intend.

REFERENCES

Adamsky, Dima, and Kjell Inge Bjerga. 2010. "Introduction to the Information-Technology Revolution in Military Affairs." *Journal of Strategic Studies* 33, no. 4: 463–468. https://doi.org/10.1080/01402390.2010.489700.

AlphaStar Team. 2019. "AlphaStar: Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning." DeepMind. 2019. https://deepmind.com/blog/article/AlphaStar-Grandmaster-level-in-StarCraft-II-using-multi-agent-reinforcement-learning.

Andrew, Christopher M. 2000. *The Mitrokhin Archive: The KGB in Europe and the West*. London: Allen Lane.

Andrew, Christopher M., and Vasili Mitrokhin. 1999. *The Sword and the Shield: The Mitrokhin Archive and the Secret History of the KGB*, 1st ed. New York: Basic Books.

Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2018. "Synthesizing Robust Adversarial Examples." *ArXiv:1707.07397 [Cs]*, June. http://arxiv.org/abs/1707.07397.

Bathaee, Yavar. 2017. "The Artificial Intelligence Black Box and the Failure of Intent and Causation." *Harvard Journal of Law and Technology (Harvard JOLT)* 31, no. 2: 889–938.

Bezdek, James C. 1992. "On the Relationship between Neural Networks, Pattern Recognition and Intelligence." *International Journal of Approximate Reasoning* 6, no. 2: 85–107. https://doi.org/10.1016/0888-613X(92)90013-P.

Biddle, Stephen. 2010. *Military Power: Explaining Victory and Defeat in Modern Battle*. Princeton: Princeton University Press.

Bilgin, Zeki. 2021. "Anomaly Localization in Model Gradients Under Backdoor Attacks against Federated Learning." *ArXiv:2111.14683 [Cs]*, November. http://arxiv.org/abs/2111.14683.

Blackstock, Paul W. 1978. *Intelligence, Espionage, Counterespionage and Covert Operations: A Guide to Information Sources*. International Relations Information Guide Series, vol. 2. Detroit: Gale Research Company.

Bose, N. K., and P. Liang. 1996. *Neural Network Fundamentals with Graphs, Algorithms, and Applications*. USA: McGraw-Hill.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*, 1st ed. Oxford: Oxford University Press.

Brooks, Rosa. 2015. "In Defense of Killer Robots." *Foreign Policy* (blog). https://foreignpolicy.com/2015/05/18/in-defense-of-killer-robots/.

Buchanan, Ben. 2020. *The Hacker and the State: Cyber Attacks and the New Normal of Geopolitics*. Cambridge, MA: Harvard University Press.

Cameron, James (director). 1984. *The Terminator*. Action / science-fiction film. Orion Pictures. 107 minutes.

Chen, Xinyun, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning." *ArXiv:1712.05526 [Cs]*, December. http://arxiv.org/abs/1712.05526.

Comiter, Marcus. 2019. "Attacking Artificial Intelligence." Belfer Center for Science and International Affairs / Harvard Kennedy School. https://www.belfercenter.org/sites/default/files/2019-08/AttackingAI/AttackingAI.pdf.

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. "Robust Physical-World Attacks on Deep Learning Visual Classification." In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1625–1634. https://doi.org/10.1109/CVPR.2018.00175.

Fischerkeller, Michael P., and Richard J. Harknett. 2017. "Deterrence Is Not a Credible Strategy for Cyberspace." *Orbis* 61, no. 3: 381–393. https://doi.org/10.1016/j.orbis.2017.05.003.

Freedberg, Sydney J., Jr. 2015. "Centaur Army: Bob Work, Robotics, and the Third Offset Strategy." *Breaking Defense* (blog). November 9. https://breakingdefense.sites.breakingmedia.com/2015/11/centaur-army-bob-work-robotics-the-third-offset-strategy/.

Gigova, Radina. 2017. “Who Putin Thinks Will Rule the World.” CNN, September 2. https://edition.cnn.com/2017/09/01/world/putin-artificial-intelligence-will-rule-world/index.html.

Gill, Indermit. 2020. “Whoever Leads in Artificial Intelligence in 2030 Will Rule the World until 2100.” *Brookings* (blog), January 17. https://www.brookings.edu/blog/future-development/2020/01/17/whoever-leads-in-artificial-intelligence-in-2030-will-rule-the-world-until-2100/.

Global Defence Technology. 2019. “Back Door for Hackers? F-35 Cyber Weaknesses in the Spotlight.” Issue no. 97, March. https://defence.nridigital.com/global_defence_technology_mar19/back_door_for_hackers_f-35_cyber_weaknesses_in_the_spotlight. Accessed December 11, 2021.

Goldfarb, Avi, and Jon R. Lindsay. 2022. “Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War.” *International Security* 46, no. 3: 7–50. https://doi.org/10.1162/isec_a_00425.

Herman, Michael. 1996. *Intelligence Power in Peace and War*. Cambridge: Cambridge University Press.

Hestness, Joel, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. “Deep Learning Scaling Is Predictable, Empirically.” *ArXiv:1712.00409 [Cs, Stat]*, December. http://arxiv.org/abs/1712.00409.

High, Peter. 2017. “Carnegie Mellon Dean of Computer Science on the Future of AI.” *Forbes*, October 30, Tech section. https://www.forbes.com/sites/peterhigh/2017/10/30/carnegie-mellon-dean-of-computer-science-on-the-future-of-ai/.

Horowitz, Michael C. 2018. “Artificial Intelligence, International Competition, and the Balance of Power (May 2018),” May. https://doi.org/10.15781/T2639KP49.

———. 2019. “When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability.” *Journal of Strategic Studies* 42, no. 6: 764–788. https://doi.org/10.1080/01402390.2019.1621174.

Keck, Catie. 2020. “How a Piece of Tape Tricked a Tesla into Reading a 35MPH Sign as 85MPH.” *Gizmodo*, February 19. https://gizmodo.com/how-a-piece-of-tape-tricked-a-tesla-into-reading-a-35mp-1841791417.

Kello, Lucas. 2013. “The Meaning of the Cyber Revolution: Perils to Theory and Statecraft.” *International Security* 38, no. 2: 7–40.

Kozma, Robert, Cesare Alippi, Yoonsuck Choe, and F. C. Morabito, eds. 2019. *Artificial Intelligence in the Age of Neural Networks and Brain Computing*. London and San Diego: Elsevier, Academic Press.

Lee, Han-Soo. 2021. “DeepBio’s AI-Based Prostate Cancer Pathology Diagnosis Wins New Tech Certification.” *KBR*, December 9. http://www.koreabiomed.com/news/articleView.html?idxno=12737.

Lei, Xinyu, Guan-Hua Tu, Alex X. Liu, Chi-Yu Li, and Tian Xie. 2018. “The Insecurity of Home Digital Voice Assistants: Vulnerabilities, Attacks and Countermeasures.” In *2018 IEEE Conference on Communications and Network Security (CNS)*, 1–9. Beijing: IEEE. https://doi.org/10.1109/CNS.2018.8433167.

Lindsay, Jon R. 2013. “Stuxnet and the Limits of Cyber Warfare.” *Security Studies* 22, no. 3: 365–404. https://doi.org/10.1080/09636412.2013.816122.

———. 2020. *Information Technology and Military Power.* Cornell Studies in Security Affairs. Ithaca, NY: Cornell University Press.

Maschmeyer, Lennart. 2021. “The Subversive Trilemma: Why Cyber Operations Fall Short of Expectations.” *International Security* 46 (2): 51–90. https://doi.org/10.1162/isec_a_00418.

Myers, Meghann. 2015. “SECNAV: F-35C Should Be Navy’s Last Manned Strike Jet.” *Navy Times*, April 16. https://www.navytimes.com/news/your-navy/2015/04/16/secnav-f-35c-should-be-navy-s-last-manned-strike-jet/.

Narodytska, Nina, and Shiva Kasiviswanathan. 2017. “Simple Black-Box Adversarial Attacks on Deep Neural Networks.” In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1310–1318. Honolulu: IEEE. https://doi.org/10.1109/CVPRW.2017.172.

NATO. 2020. “NATO 2030: United for a New Era.” https://www.nato.int/nato_static_fl2014/assets/pdf/2020/12/pdf/201201-Reflection-Group-Final-Report-Uni.pdf.

Nguyen, Anh, Jason Yosinski, and Jeff Clune. 2015. “Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images.” *ArXiv:1412.1897 [Cs]*, April. http://arxiv.org/abs/1412.1897.

Payne, Kenneth. 2018. “Artificial Intelligence: A Revolution in Strategic Affairs?” *Survival* 60, no. 5: 7–32. https://doi.org/10.1080/00396338.2018.1518374.

Pietrucha, Mike. 2015. “Why the Next Fighter Will Be Manned, and the One after That.” *War on the Rocks* (blog), August 5, 2015. https://warontherocks.com/2015/08/why-the-next-fighter-will-be-manned-and-the-one-after-that/.

Qi, Xiangyu, Tinghao Xie, Ruizhe Pan, Jifeng Zhu, Yong Yang, and Kai Bu. 2021. “Towards Practical Deployment-Stage Backdoor Attack on Deep Neural Networks.” *ArXiv:2111.12965 [Cs]*, November. http://arxiv.org/abs/2111.12965.

Roff, Heather M., and David Danks. 2018. “‘Trust but Verify’: The Difficulty of Trusting Autonomous Weapons Systems.” *Journal of Military Ethics* 17, no. 1: 2–20. https://doi.org/10.1080/15027570.2018.1481907.

Sayler, Kelley M. 2020. “Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems.” December, 3. https://sgp.fas.org/crs/natsec/IF11150.pdf

Scharre, Paul. 2018a. “A Million Mistakes a Second.” *Foreign Policy* (blog). 2018. https://foreignpolicy.com/2018/09/12/a-million-mistakes-a-second-future-of-war/.

———. 2018b. *Army of None: Autonomous Weapons and the Future of War*, 1st ed. New York and London: W. W. Norton and Company.

ScienceDaily. 2021. “New Imaging Technology May Reduce Need for Skin Biopsies.” November 18. https://www.sciencedaily.com/releases/2021/11/211118203054.htm.

Sheu, Yi-han. 2020. “Illuminating the Black Box: Interpreting Deep Neural Network Models for Psychiatric Research.” *Frontiers in Psychiatry* 11: 1091. https://doi.org/10.3389/fpsyt.2020.551299.

“Stopping Killer Robots.” 2020. Human Rights Watch. August 10. https://www.hrw.org/report/2020/08/10/stopping-killer-robots/country-positions-banning-fully-autonomous-weapons-and.

Towards AI. 2021. “Best Public Datasets for Machine Learning and Data Science.” Medium. April 2. https://pub.towardsai.net/best-datasets-for-machine-learning-data-science-computer-vision-nlp-ai-c9541058cf4f.

Tyukin, Ivan Y., Desmond J. Higham, Eliyas Woldegeorgis, and Alexander N. Gorban. 2021. “The Feasibility and Inevitability of Stealth Attacks.” *ArXiv:2106.13997 [Cs]*, October. http://arxiv.org/abs/2106.13997.

Umbrello, Steven, Phil Torres, and Angelo F. De Bellis. 2020. “The Future of War: Could Lethal Autonomous Weapons Make Conflict More Ethical?” *AI and Society* 35, no. 1: 273–282. https://doi.org/10.1007/s00146-019-00879-x.

US Department of Defense. 2017. “Autonomy in Weapons Systems.” Directive 3000.09. https://irp.fas.org/doddir/dod/d3000_09.pdf.

Vazquez, Annemarie. 2019. “Laws and Lawyers: Lethal Autonomous Weapons Bring LOAC Issues to the Design Table, and Judge Advocates Need to Be There.” *Military Law Review* 228, no. 1: 89–131.

Whyte, Christopher. 2020. “Problems of Poison: New Paradigms and ‘Agreed’ Competition in the Era of AI-Enabled Cyber Operations.” In *20/20 Vision: The Next Decade*, eds. T. Jančárková, L. Lindström, M. Signoretti, I. Tolga, and G. Visky. 12th International Conference on Cyber Conflict. Tallinn: NATO CCDCOE Publications. https://ccdcoe.org/uploads/2020/05/CyCon_2020_12_Whyte.pdf.

Zhang, Feng, Huibing Fan, Keju Wang, Yongjin Zhao, Xiaoxi Zhang, and Yang Ma. 2021. “Research on Intelligent Target Recognition Integrated With Knowledge.” *IEEE Access* 9: 137107–137115. https://doi.org/10.1109/ACCESS.2021.3116866.

Zhou, Wenying, Yang Yang, Cheng Yu, Juxian Liu, Xingxing Duan, Zongjie Weng, Dan Chen et al. 2021. “Ensembled Deep Learning Model Outperforms Human Experts in Diagnosing Biliary Atresia from Sonographic Gallbladder Images.” *Nature Communications* 12, no. 1: 1259. https://doi.org/10.1038/s41467-021-21466-z.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

'Responsibility to Detect?': Autonomous Threat Detection and its Implications for Due Diligence in Cyberspace

Arun Mohan Sukumar
PhD Candidate
The Fletcher School of Law and Diplomacy
Tufts University
Medford, MA, United States
arun.sukumar@tufts.edu

Abstract: Private and public organizations have long relied on intrusion detection systems to alert them of malicious activity in their digital networks. These systems were designed to detect threat signatures in static networks or infer anomalous activity based on their security 'logs'. They are, however, of limited use to detect threats across heterogeneous, modern-day networks, where computing resources are distributed across cloud or routing services. Recent advancements in machine learning (ML) have led to the development of autonomous threat detection (ATD) applications that monitor, evaluate, and respond to malicious activity with minimal human intervention. The use of 'intelligent' and programmable algorithms for ATD will reduce incident response times and enhance the capacity of states to detect threats originating from any layer of their territorial information and communications technologies (ICT) infrastructure. This paper argues that ATD technologies will influence the evolution of a due diligence rule for cyberspace by raising the standard of care owed by states to prevent their networks from being used for malicious, transboundary ICT activities. This paper comprises five sections. Section 1 introduces the paper and its central argument. Section 2 outlines broad trends and operational factors pushing public and private entities towards the adoption of ATD. Section 3 offers an overview of a typical ATD application. Section 4 analyses the impact of ATD on the due diligence obligations of states. Section 5 presents the paper's conclusions.

Keywords: *due diligence, autonomous systems, international law, threat detection, machine learning*

1. INTRODUCTION

The use of ATD technologies – applications capable of identifying, analysing, and in some cases, even responding to malicious activity in digital networks with minimal or no human supervision – has been contemplated for nearly three decades.[1] Today, with advancements in computing, it is possible to implement at scale ATD that relies on artificial intelligence (AI) / machine learning (ML) models. If computing advancements have made the widespread deployment of ATD possible, the pervasive digitalization of services and 'things' has made autonomous detection somewhat necessary. The topology of the modern digital network is extremely diverse, consisting of software and hardware whose ownership or management is often shared across vendors and geographies. In particular, the shift towards 'work-from-home', precipitated overnight by the COVID-19 pandemic, has resulted in businesses relying on cloud and network services that are scattered regionally and even globally. The terms 'network operator' and 'sysadmin' (short for 'system administrator') – the individual or organization responsible, inter alia, for managing the security of their enterprise's virtual infrastructure – are themselves misnomers in the contemporary era, given that they have limited visibility over vulnerabilities or attack vectors across network components. The traditional notion of a 'security perimeter', understood as the outer limits of a spatially and physically bound intranet, has suddenly become outdated.[2] Yet, threat detection tools have been slow to respond to this shift. Repetto et al. argue that cyber security applications are currently deployed in 'vertical silos' across a network, protecting the cloud service, applications, devices, enterprise architecture, etc.[3] This is not a viable solution to protect heterogeneous environments and often leads to malicious actors seeking out the path of least resistance in a network. ATD applications autonomously interface with various components of a digital network, drawing relevant and contextual information from those components to learn and detect potentially malicious threats. Consequently, these applications provide human operators with greater visibility over their fragmented network and obviate the cumbersome, manual monitoring of threats across various network components.

The roll-out of ATD in the public and private sector will shape not only cyber security practices but also the application of international law to state behaviour in cyberspace. This paper argues that the ability of ATD applications to detect and notify network operators of malicious activity will raise the standard of care owed by states to address significant transboundary harm in cyberspace. In other words, the adoption of ATD

1 See generally, H Debar, M Becker, and D Siboni, 'A Neural Network Component for an Intrusion Detection System' in *Proceedings of the 1992 IEEE Computer Society Symposium on Research in Security and Privacy*, 1992, 240–250; Jeremy Frank, 'Artificial Intelligence and Intrusion Detection: Current and Future Directions', in *Proceedings of the 17th National Computer Security Conference*, 1994, 1–12.

2 R. Rapuzzi and M. Repetto, 'Building Situational Awareness for Network Threats in Fog/Edge Computing: Emerging Paradigms beyond the Security Perimeter Model', *Future Generation Computer Systems* 85 (August 1, 2018): 235.

3 Matteo Repetto et al., 'An Autonomous Cybersecurity Framework for Next-Generation Digital Service Chains' (2021) 4 *Journal of Network and Systems Management* 29, 36.

will decisively influence the cyber due diligence principle. This paper comprises five sections. Section 2 outlines broad trends and operational factors pushing public and private entities towards adopting ATD to monitor their networks. Section 3 offers an overview of a typical ATD framework. Section 4 analyses the impact of ATD on the due diligence obligations of states. Section 5 presents the paper's conclusions.

2. THREAT DETECTION IN FRAGMENTED NETWORKS

The 'virtualization' of computing functions has lent incredible complexity to the modern-day digital network. As Valenza et al. note, there is no longer a linear connection that can be drawn between the digital application and the end-user device:[4] if its data is processed in cloud or edge servers, its networking functions are also outsourced for reasons of efficiency and costs. The fragmentation of the digital network makes it difficult for any one operator to monitor the entire network's security. This section reviews some of the factors identified in the technical literature that make conventional threat detection challenging in the contemporary era.

A. Multi-tenancy

Multi-tenancy refers to the sharing of software and hardware resources by a group of entities. These resources usually take the form of servers or databases used for storing or processing data. Cloud computing is the most common example of a multi-tenant framework where an umbrella vendor, such as Amazon Web Services (AWS) or Microsoft Azure, hosts several clients on its servers. In a multi-tenant environment, the 'tenants' manage their own end-point protocols, potentially creating multiple vectors of vulnerability for the whole infrastructure.[5] Tenants are also heavily dependent on the resources of the host, creating a central point of cyber security failure.[6] The challenges in securing multi-tenant environments are not limited to commercial networks but also extend to industrial control systems and critical infrastructure (CI). Indeed, the dependency of CI providers on cloud-based services has elicited stringent policy measures from states. For example, in December 2021, Australia passed legislation allowing government agencies to commandeer the resources of 'critical infrastructure assets' – including private cloud operators – to respond to a serious cyber attack in

4 Fulvio Valenza, Matteo Repetto and Stavros Shiaeles, 'Guest Editorial: Special Issue on Novel Cyber-Security Paradigms for Software-Defined and Virtualized Systems' *Computer Networks* 193 (July 5, 2021): 108126.

5 Repetto et al. (n 3) 37.

6 Wayne J Brown, Vince Anderson and Qing Tan, 'Multitenancy – Security Risks and Countermeasures' in *2012 15th International Conference on Network-Based Information Systems*, 2012, 7–13; see generally, 'SolarWinds Breach Exposes Hybrid Multicloud Security Weaknesses' (VentureBeat, 16 May 2021) <https://venturebeat.com/2021/05/16/solarwinds-breach-exposes-hybrid-multi-cloud-security-weaknesses/> accessed 10 November 2021.

exigent circumstances.[7] Such *post facto* measures, however, still do not address the challenge of timely identification of cyber attacks on multi-tenant environments.

In addition to outsourcing data storage and processing, public and private entities also delegate networking functions to third parties. This delegation, called Network Functions Virtualization (NFV), involves the handing over of packet-switching and routing functions, and even network-associated services such as firewall security, to an external vendor, such as Cloudflare, Akamai, or VMWare. NFV allows small organizations to conserve costs associated with purchasing and maintaining networking infrastructure, but the diversity of hardware components also makes it difficult to 'isolate and contain malware' within their networks.[8] As Firoozjaei et al. note, NFV exposes networks not only to infrastructure-based threats but also to targeted end-user threats because network services such as firewalls or secure sockets layer (SSL) gateways give NFV providers 'complete dominance over the user's information'.[9] Internet of Things (IoT) systems in particular have come to depend on NFV, given the limited computing power of IoT devices and the requirement for low latency of data traffic in some cases.[10]

Multi-tenant services have limited incentives to guarantee the security of their clients. As Schneier and Herr note, 'security is largely an externality' for cloud vendors because the cost of cyber attacks is borne by users and client organizations.[11] With CI, those costs are often borne by governments. Consequently, cyber attacks in multi-tenant environments are likely to persist, making timely identification all the more relevant.

B. Widespread Use of Application Programming Interfaces

Partly on account of the rise of multi-tenancy, but owing primarily to the explosive growth of the platform economy and social media, the internet is witnessing unprecedented 'API-fication'. Application programming interfaces (APIs) are lines of code that allow software to communicate with each other. They are digital railroads, built either by governments or private actors, that allow third parties to retrieve user data, integrate with multisided platforms, and in the case of NFV, communicate with

7 'Government Assistance' (Australian Government Department of Home Affairs) <www.homeaffairs.gov.au/about-us/our-portfolios/national-security/security-coordination/security-of-critical-infrastructure-act-2018-amendments/government-assistance> accessed 9 December 2021; 'Security Legislation Amendment (Critical Infrastructure) Act 2021', No. 124, 2021 (The Parliament of the Commonwealth of Australia), 63-74, <https://www.legislation.gov.au/Details/C2021A00124> accessed 10 November 2021.

8 'What Is Network Functions Virtualization (NFV)? | VMware Glossary' (VMware) <www.vmware.com/topics/glossary/content/network-functions-virtualization-nfv.html> accessed 9 January 2022.

9 Mahdi Daghmehchi Firoozjaei et al., 'Security Challenges with Network Functions Virtualization' (2017) 67 *Future Generation Computer Systems*, 315, 320.

10 See Nikos Bizanis and Fernando A Kuipers, 'SDN and Virtualization Solutions for the Internet of Things: A Survey' (2016) 4 *IEEE Access* 5591–5606.

11 Bruce Schneier and Trey Herr, 'Russia's Hacking Success Shows How Vulnerable the Cloud Is' (*Foreign Policy*, 24 May 2021) <https://foreignpolicy.com/2021/05/24/cybersecurity-cyberattack-russia-hackers-cloud-sunburst-microsoft-office-365-data-leak/> accessed 10 November 2021

applications as well as routing infrastructure in order to direct network traffic.[12] APIs play an important role in ensuring the interoperability of digital networks and seamless delivery of digital services. API 'calls' make up 83% of online traffic today.[13] With the proliferation of APIs, however, have emerged attendant risks. While acting as the internet's connective tissue, APIs also considerably expand the cyber attack surface.[14] Specifically, APIs expose networks to data breaches (the most common API security incidents[15]), person-in-the-middle attacks, and Distributed Denial of Service (DDoS) attacks that disrupt the availability of services, among others.[16]

Despite these concerns, API security has largely been sidestepped in favour of ease of adoption.[17] Weak authentication mechanisms, data leakage, and poor auditing of APIs have already led to major cyber security incidents and pose a challenge for businesses and policymakers alike.[18] The diversity and differential security policies of APIs, especially in fragmented networks, make threat detection extremely difficult.

C. Uneven Cyber Security Policy Landscape

Despite vulnerabilities posed by multi-tenant frameworks and the widespread use of APIs, most states have traditionally equated cyber security with data protection at the 'last mile'.[19] National regulatory instruments often focus exclusively on the

12 Truman Boyes et al., 'Accelerating NFV Delivery with OpenStack: Global Telecoms Align Around Open Source Networking Future' (OpenStack Foundation Report, 2016) <https://object-storage-ca-ymq-1.vexxhost.net/swift/v1/6e4619c416ff4bd19e1c087f27a43eea/www-assets-prod/marketing/OpenStack-NFV-Print.pdf> accessed 3 November 2021.

13 Akamai, 'State of the Internet / Security: Retail Attacks and API Traffic Report' (2019) <https://www.akamai.com/content/dam/site/it/documents/state-of-the-internet/state-of-the-internet-security-retail-attacks-and-api-traffic-report-2019.pdf> accessed 3 November 2021.

14 'Akamai: API: The Attack Surface That Connects Us All' (2021) 11 *Computer Fraud & Security* 4.

15 Brian Krebs, 'USPS Site Exposed Data on 60 Million Users – Krebs on Security' (*KrebsonSecurity* 21 November 2018) <https://krebsonsecurity.com/2018/11/usps-site-exposed-data-on-60-million-users/> accessed 3 November 2021; Dan Salmon, 'I Scraped Millions of Venmo Payments. Your Data Is at Risk' (*Wired*, 26 June 2019) <www.wired.com/story/i-scraped-millions-of-venmo-payments-your-data-is-at-risk/> accessed 3 November 2021; 'Rapid Growth of APIs Has Led to Security Risks for the Enterprise' (Cloudflare) <https://www.cloudflare.com/insights-api-proliferation/> accessed 9 January 2022.

16 Torsten George, 'The Next Big Cyber-Attack Vector: APIs' (*SecurityWeek*, 28 June 2018) <www.securityweek.com/next-big-cyber-attack-vector-apis> accessed 3 November 2021; 'API Attacks Are Both Underdetected and Underreported' (*Help Net Security*, 28 October 2021) <www.helpnetsecurity.com/2021/10/28/security-concerns-api/> accessed 3 November 2021.

17 See 'API Data Breaches in 2020' 9*CloudVector*, 23 December 2020) <www.cloudvector.com/api-data-breaches-in-2020/> accessed 5 November 2021.

18 Lindsey O'Donnell, 'Microsoft OAuth Flaw Opens Azure Accounts to Takeover' (*Threatpost*, 2 December 2019) <https://threatpost.com/microsoft-oauth-flaw-azure-takeover/150737/> accessed 9 November 2021]; 'Uber Disclosed on HackerOne: Sensitive User Information Disclosure at Bonjour.Uber.Com/Marketplace/_rpc via the "userUuid" Parameter' (HackerOne) <https://hackerone.com/reports/542340> accessed 9 January 2022; 'Amazon's Ring Neighbors App Exposed Users' Precise Locations and Home Addresses' (*TechCrunch*) <https://social.techcrunch.com/2021/01/14/ring-neighbors-exposed-locations-addresses/> accessed 9 January 2022; 'Information Leakage in AWS Resource-Based Policy APIs' (*Unit42*, 17 November 2020) <https://unit42.paloaltonetworks.com/aws-resource-based-policy-apis/> accessed 10 November 2021.

19 Jeff Kosseff, 'Defining Cybersecurity Law' (2018) 103 Iowa L. Rev. 985, 995.To be sure, this has changed in recent years. See generally, Agnes Kasper and Alexander Antonov, *Towards Conceptualizing EU Cybersecurity Law*, Discussion Paper / Zentrum Für Europäische Integrationsforschung, C 253 (Bonn: Zentrum für Europäische Integrationsforschung, Rheinische Friedrich-Wilhelms Universität, 2019) 26.

relationship between the end-user and their device or application, laying down broad guidelines on the types of personal and non-personal data that private and public entities can collect and store. Other network layers and infrastructure are often ignored. As a result, outside of data breaches, API development or the roles and responsibilities of cloud and NFV service providers have been poorly regulated in most jurisdictions. The lack of a clear regulatory framework on this issue makes vulnerabilities and incident reporting largely a factor of market practices, which are hardly uniform within and across states.

3. ATD: THE FUTURE OF THREAT DETECTION

Confronted thus by a fragmented network environment and uneven policy standards, states and private actors may currently pursue three options for whole-of-network threat detection.

Layered security: Organizations can depend on separate services to monitor and protect their network infrastructure, cloud-based resources, applications, and terminal devices. Given the difficulty and costs involved in integrating threat inputs from different sources, this approach is unlikely to be preferred by most network administrators.

Host-based security: Service providers such as AWS, Alibaba Cloud, and Cloudflare have begun offering services that monitor network traffic, perform authentication, and track API security.[20] Most of these services allow network administrators to 'remotely manage'[21] incidents from a centralized console.

Third-party security: Network operators can also rely on the services of a third party, which is usually a commercial entity (for example, Checkpoint, CrowdStrike, Mandiant). Third-party security offers flexibility to organizations that may want to avoid a lock-in of their threat detection capabilities with the host that provides multi-tenant services. Additionally, many cyber security vendors claim to offer multi-cloud threat detection, allowing clients to identify anomalous behaviour across various services.[22]

20 'Security at the Edge: Core Principles' (AWS, 24 September 2021) <https://d1.awsstatic.com/whitepapers/Security/security-at-the-edge.pdf> accessed 3 November 2021; 'Getting Started with Secure Access Service Edge: A Guide to Secure and Streamline Your Network Infrastructure' (Cloudflare, 22 October 2021) <www.cloudflare.com/static/52527ba193cc7ab0da6c23075d093ab3/Cloudflare_One_SASE_Whitepaper.pdf> accessed 3 November 2021; 'Alibaba Cloud Security Services' (Alibaba Cloud) <www.alibabacloud.com/product/security> accessed 9 January 2022.

21 See 'AWS IoT Device Management Features – AWS' (Amazon Web Services, Inc.) <https://aws.amazon.com/iot-device-management/features/> accessed 9 January 2022.

22 'Cloud Native Security – Security Automated Everywhere' (Checkpoint, 2021), <www.checkpoint.com/downloads/products/cloudguard-cloud-native-security-datasheet.pdf> accessed 3 November 2021.

In the wake of the COVID-19 pandemic, more organizations have sought 'full-stack observability'[23] over disparate components of their network. They may move towards host-based and third-party security offerings, as described above. These services have, in turn, begun implementing AI/ML models to perform intrusion detection.[24] AI/ML-based detection turns the problem of multi-tenancy into a solution. As with other aspects of digital networking, AI-based security functions are also increasingly offloaded to cloud or edge servers. With more computing resources available to process large volumes of traffic, it is today possible to train algorithms 'remotely' to detect threats and respond to them with low latency. This is especially useful in the case of IoT networks, which have increasingly been targets of DDoS attacks.[25]

However, the use of AI/ML models to detect cyber security threats has hitherto tended to fall into two categories: algorithms that can detect anomalies in 'static' topologies, that is, those networks where routing is predictable and where ports of entry and exit remain constant, or algorithms that learn to detect very specific malware in a network, whether it is based on unusual signatures or traffic patterns.[26] In dynamic multi-tenant environments, such applications are of limited use.[27] Nevertheless, newer applications of ATD leverage advancements in edge and cloud computing to obtain greater visibility over heterogeneous network environments. Such applications lean on a centralized architecture that performs whole-of-network monitoring, irrespective of the changing elements of its infrastructure or applications. The following paragraphs review the functioning of a typical ATD application.

In simple terms, ATD applications fetch information from various components of the network into 'clean rooms' that subsequently process such data to identify threats. ASTRID, a multistakeholder pilot project supported by the European Union, offers an example of such an ATD application.[28] ASTRID – which stands for AddreSsing ThReats for virtualIzeD services – provides a conceptual and technical framework to 'decouple' security functions from the overall functioning of individual network components. It does so by creating a 'centralized architecture' that collects 'security information, data, and events' from various network sources. This architecture

23 Erwan Paccard, 'Why Full-Stack Observability Is Critical for a Successful DevSecOps Approach' (*Computing*, 15 December 2021) <www.computing.co.uk/sponsored/4042095/stack-observability-critical-successful-devsecops-approach> accessed 12 November 2021.

24 'Endpoint Protection Software Explained' (Crowdstrike) <www.crowdstrike.com/cybersecurity-101/endpoint-security/endpoint-protection-software/> accessed 9 January 2022; 'Endpoint Protection Buyer's Guide 2020' (Checkpoint) <https://app.hushly.com/runtime/content/JGq2xOVJoBplBawJ> accessed 12 November 2021; 'Amazon Detective – AWS' <https://aws.amazon.com/detective/?c=sc&sec=srv> accessed 9 January 2022.

25 Liang Xiao et al., 'IoT Security Techniques Based on Machine Learning: How Do IoT Devices Use AI to Enhance Security?' 2018 35(5) *IEEE Signal Processing Magazine*, 41–49.

26 Maruthi Rohit Ayyagari et al., 'Intrusion Detection Techniques in Network Environment: A Systematic Review' (2021) 27(2) *Wireless Networks*, 1269.

27 Daniel Spiekermann and Jörg Keller, 'Unsupervised Packet-Based Anomaly Detection in Virtual Networks' (2021) 192 *Networks* 2.

28 'ASTRID Project: A Cybersecurity Framework for Virtualized Services' <https://www.astrid-project.eu/project.html> accessed 9 January 2022.

comprises a data plane, control plane, and management plane.[29] The data plane is the programmable component of the framework that collects and maintains security-related logs, events, and traffic metrics spanning the network. Most components of the network have event- and log-reporting capabilities built into their software, and the data plane relies on 'lightweight hooks', that is, APIs, to query and retrieve security information from their kernels or libraries. The control plane is a collection of ML algorithms that retrieve information from the data plane, and through it, obtains 'complete visibility' over the network. These algorithms evaluate the security information and identify threats based on anomalous behaviour. Finally, the management plane is the human-facing element of this architecture, which communicates threats in real time to network administrators and suggests remedial measures.

The operationalization of such an ATD architecture will depend on two technical factors. First, it requires the availability of adequate computing resources to perform 'clean room' functions. The data plane does not simply collect logs and events from network components but also dynamically adjusts the scope and frequency of reporting as necessitated by circumstances.[30] If potentially malicious behaviour is identified in one section of the network – for example, suspicious API calls or unusual router volumes – then the ATD application channels greater detection and remedial resources towards it. Similarly, if the network relies on a new cloud or NFV service, the ATD application may seek more data from it to learn its behaviour and train its algorithms. Such adjustments necessitate adequate computing resources. Second, ML-driven threat detection through a 'command-and-control' architecture requires interfacing with security functions of other network components (for example, firewalls, packet inspection tools, other analytics software, etc.). This is only possible if the various network services adopt common and interoperable interfaces allowing the data plane to ping and access relevant reporting information.

Both technical factors are close to realization today. As already noted, advancements in edge and cloud computing enhance the programmability of ATD applications. With respect to common security interfaces, there has been a notable parallel effort from within the technical community to develop interoperable standards that monitor heterogeneous networks. Since 2014, a Birds of a Feather (BoF) group in the Internet Engineering Task Force (IETF) has sought to promote discussion on a common 'Interface 2 Network Security Functions' (I2NSF).[31] This group, which includes volunteers from prominent global technology companies, has sought 'a standardized interface to control and monitor the rule sets that network security functions [NSFs]

29 ASTRID Consortium, 'D1.2 – ASTRID Architecture' 31–33 <https://cyberwatching.eu/sites/default/files/D1.2%20-%20ASTRID%20architecture.pdf> accessed 12 November 2021.

30 R Bolla, A Carrega, and M Repetto, 'An Abstraction Layer for Cybersecurity Context' in *2019 International Conference on Computing, Networking and Communications (ICNC)* (2019) 215.

31 'RFC 8192 Interface to Network Security Functions (I2NSF): Problem Statement and Use Cases' (IETF Datatracker) <https://datatracker.ietf.org/doc/rfc8192/> accessed 9 January 2022.

use to treat packets traversing through these NSFs'.[32] In other words, I2NSF aims to create 'vendor-agnostic' protocols that allow for a seamless flow of threats-related information – whether through centralized or distributed architecture – among various network components by accessing their security functions and capabilities. The BoF group has specifically emphasized the need for an interface sensitive to periodic updates of security policies or configurations by stand-alone network services, which is crucial to an 'autonomous security system'.[33] Indeed, given the synergies between both endeavours, the ASTRID project highlights in detail the characteristics of the I2NSF proposal.[34] Whether or not this specific IETF initiative succeeds,[35] it is only reasonable to conclude similar efforts – including those by the private sector[36] – will mushroom in the coming years.

ATD frameworks such as ASTRID enhance the visibility of system administrators over their networks and, through their programmability, also offer the system administrators greater control when addressing threats and vulnerabilities unique to their organizations. Beyond market consequences, the policy impact of ATD is equally notable. For instance, the 'command-and-control' model of ATD applications allows states to detect and respond to cyber security threats to publicly owned CI, even if such CI relies on private cloud/NFV services or even if those services are located in another country. ATD could also provide states with accurate and instantaneous knowledge of transboundary malicious activity emanating from or transiting through their territory. The following section addresses this possibility in greater detail.

How exactly could states rely on ATD frameworks? Some states may develop a 'plug-and-play' ATD application, using its technical framework to enforce cyber security policies on monitoring networks for harmful activity, and require all private and public operators based in its territory to adopt it. Other states could develop APIs for their Computer Security Incident Response Teams (CSIRTs) that interface with private ATD applications. As a result, CSIRTs would be automatically notified whenever potentially malicious threats are detected by those applications.

32 S Hares et al., 'Interface to Network Security Functions (I2NSF): Problem Statement and Use Cases' (RFC Editor, July 2017) 7 <www.rfc-editor.org/rfc/pdfrfc/rfc8192.txt.pdf> accessed 12 November 2021

33 'Re: [I2nsf] I2NSF Re-Chartering Text' <https://mailarchive.ietf.org/arch/msg/i2nsf/rn1F7BSqqEzI15ApV2c0UHjmbz8/> accessed 9 January 2022.

34 ASTRID Consortium (n 29) 68–72.

35 For an overview of the Internet Engineering Steering Group's competing views on the proposal, see 'Ballot for Draft IETF RFC (Interface to Network Security Functions (I2NSF): Problem Statement and Use Cases)' (IETF Datatracker) https://datatracker.ietf.org/doc/rfc8192/ballot/> accessed 9 January 2022.

36 Jordan Novet, 'Amazon's Outage and HashiCorp's IPO Point to a Future with Multiple Clouds' (CNBC, 12 December 12, 2021), https:// <www.cnbc.com/2021/12/12/aws-outage-and-hashicorp-ipo-point-to-a-multicloud-future.html> accessed 9 November 2021.

4. ATD AND CYBER DUE DILIGENCE

By enhancing their ability to detect harmful cyber activity, ATD applications could also influence the scope of duties states have with respect to preventing their territory from being used to launch or relay cyber operations targeting another state. The due diligence principle in international law, as enunciated by the International Court of Justice in *The Corfu Channel Case*, refers to the obligation of a state 'not to allow knowingly its territory to be used for acts contrary to the rights of other States.'[37] The due diligence principle requires that states take all reasonable steps necessary to prevent and mitigate activities on their territory that could cause 'significant transboundary harm'.[38] The standard of care owed by states to prevent and stop harmful transboundary activities is proportionate to the risk involved in such activities. As this formulation suggests, the due diligence principle essentializes an obligation of conduct, and not an obligation of result, that is, one determined by the outcome of a state's efforts to prevent transboundary harm.[39] Nevertheless, two important considerations attach themselves to any domain-specific due diligence rule. The first, as recognized by the arbitration tribunal in the *Alabama* case,[40] is that the standard of care owed by states is not the same as that which they 'ordinary employ in their domestic concerns'.[41] The requirement of due diligence stems from the 'international duties' of states, and as such, it is not enough to treat activities causing transboundary harm in the same manner as those whose effects are territorial. Second, the International Law Commission (ILC) has noted that the due diligence principle also requires states to 'keep abreast of technological changes'.[42] The ILC's commentary to the 2001 Draft Articles on Prevention of Transboundary Harm from Hazardous Activities emphasizes perceptions of 'reasonable' or 'appropriate' measures to prevent transboundary harm may evolve because of advancements in science and technology.[43]

Both elements of the due diligence principle are relevant in the context of cyber security. A cyber-specific due diligence rule, if one exists, would impose on states an obligation to monitor and prevent cyber operations that cause significant

37 *Corfu Channel Case (United Kingdom v Albania)* (Judgment of 9 April) [1949] ICJ Rep 22.

38 International Law Commission, *Draft Articles on Prevention of Transboundary Harm from Hazardous Activities, with Commentaries*, Article 3, (2001) UN Doc. A/56/10.

39 Antal Berkes, 'The Standard of "Due Diligence" as a Result of Interchange between the Law of Armed Conflict and General International Law' (2018) 23(3) *Journal of Conflict and Security Law* 433–460; Timo Koivurova, 'Due Diligence' (last updated February 2010), in A Peters and R Wolfrum (eds), *The Max Planck Encyclopedia of Public International Law* (Oxford University Press 2008–), < https://opil.ouplaw.com/view/10.1093/law:epil/9780199231690/law-9780199231690-e1034?rskey=Nvr0gA&result=1&prd=MPIL> accessed 3 March 2022; cf Antonio Coco and Talita de Souza Dias, '"Cyber Due Diligence": A Patchwork of Protective Obligations in International Law' (2021) 32(3) *European Journal of International Law* 773.

40 Alabama Claims Arbitration (1872) 1 Moore Intl Arbitrations 495.

41 Richard Mackenzie-Gray Scott, 'Due Diligence as a Secondary Rule of General International Law' (2021) 34(2) *Leiden Journal of International Law* 343, 351.

42 International Law Commission, *Draft Articles on Prevention of Transboundary Harm from Hazardous Activities, with Commentaries*, 154 (2001) UN Doc. A/56/10.

43 ibid.

transboundary harm. This duty of care would extend beyond measures to address domestic cyber crime and include steps taken specifically to address transboundary ICT activities. Following the commentary to the ILC Draft Articles, a 'cyber' due diligence rule would also require that states progressively adopt new technologies to monitor and mitigate harmful transboundary activity on their networks. Needless to say, these considerations place strong positive obligations upon states to prevent their territory from being used for harmful cross-border activities. Partly because of the difficulty in implementing those obligations, states do not all agree (as of March 2022) on the existence of a due diligence rule for cyberspace.[44] Even those legal scholars who acknowledge the existence of a cyber due diligence principle 'recognize a more limited duty' than that applicable to other domains – namely, a duty only to stop cyber operations and not to 'prevent, or even monitor' them.[45]

Despite ambiguity as to the precise scope of a cyber due diligence principle, there appears to be growing consensus among states that they should not 'knowingly allow their territory to be used for internationally wrongful acts' using ICTs.[46] This unique formulation originally appeared in the consensus report of the 2015 UN Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (GGE), which can be considered the lodestar for non-binding and voluntary guidelines on state behaviour in cyberspace. The 2019–2021 UN GGE built on the 2015 report's norms and identified an expectation on states to take all 'appropriate, reasonably available, and feasible steps to… detect, investigate, and address' internationally wrongful acts emanating from or transiting through their territory, provided they are 'aware or notified in good faith' of such acts.[47] The norms articulated by the GGE are not binding, but they represent a clear expression of intent on the part of states to move towards a due diligence regime on cyberspace. This argument is further strengthened by the fact that the 'zero' and 'first' drafts of the 2019–2021 Open-Ended Working Group (OEWG) on ICT security declared states should 'ensure that their territory is not used by non-state actors acting on the instruction or under the control of a state to commit [internationally wrongful] acts'.[48] The language of the drafts reflected an attempt to align the due diligence requirement with that of state responsibility and restrict the scope of positive obligations only to those instances where cyber operations were

44 For an overview of national positions of key states on due diligence, see 'Due Diligence – International Cyber Law: Interactive Toolkit' (Cyber Law Toolkit) <https://cyberlaw.ccdcoe.org/wiki/Due_diligence> accessed 9 January 2022.

45 ibid.

46 'Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security', UN Doc. A/70/174 (2015), 8/17.

47 'Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security', UN Doc. A/76/135 (2021), 10/26.

48 Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security (2021) 'Draft substantive report (zero draft)', A/AC.290/2021/L.2, 6/18. Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security (2021) 'Substantive Report [FIRST DRAFT]', 14, <https://front.un-arm.org/wp-content/uploads/2021/03/210301-First-Draft.pdf> accessed 6 December 2021.

attributable to the state. This formulation was opposed on the same ground in the final negotiating session of OEWG in February 2021[49] and consequently moved to the Chair's Summary for lack of consensus, whereas the more expansive GGE formulation noted above was later adopted in June 2021. The rejection of the OEWG draft language arguably indicates a desire on the part of most states to develop a stand-alone, *sui generis* framework on cyber due diligence.

Critical to the realization and, indeed, effectiveness of a due diligence regime is conceptual clarity on what it means for a state to have 'knowledge' of harmful, transboundary cyber operations. Knowledge is not only a 'constitutive element' of due diligence,[50] but also an important technical consideration. However, there is currently no congruence between the legal and the technical thresholds of 'knowledge' required to operationalize the due diligence obligation. The GGE's guidance suggests states should be 'aware or notified in good faith' of such activity to trigger their due diligence obligations.[51] A state may legally be considered 'aware' of transboundary malicious activity based on its actual or constructive knowledge ('should have known') of the activity.[52] A state may have actual knowledge if it receives 'credible information that a harmful cyber operation is underway from its territory'.[53] In technical terms, however, the compromising of digital infrastructure – as with command-and-control servers and attack surfaces in the case of botnet attacks[54] – begins well before the attack commences. At this stage, both the motives of the attacker and their intended target are usually unclear.[55] In other words, the legal threshold of 'actual knowledge' of the originator or transit state is often too high to pursue a meaningful implementation of the due diligence principle. Once an attack has commenced, it may in fact be challenging for a state to terminate the activity without avoiding serious disruptions to its domestic digital infrastructure or services. In other words, a state with 'actual knowledge' of malicious activity could legitimately claim inability to exercise its due diligence obligations on the ground that the termination of such activity demands an unreasonable technical effort on its part. On the other hand, the determination of a state's 'constructive knowledge' about the transboundary effects of malicious activity solely on the basis of anomalous behaviour on its digital networks is technically difficult

49 See generally, 'Comments by Germany on the OEWG Zero Draft Report', 2, < https://front.un-arm.org/wp-content/uploads/2021/02/Germany-Written-Contribution-OEWG-Zero-Draft-Report_clean.pdf> accessed 4 January 2022 ; 'The Netherlands – Written Proposals to OEWG Zero Draft', 2, < https://front.un-arm.org/wp-content/uploads/2021/02/Netherlands-OEWG-informals-intervention-Feb-2021.pdf> accessed 4 January 2022.

50 Michael N Schmitt (ed), 'Due Diligence', in *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (2nd edn, Cambridge: Cambridge University Press 2017), Rule 6, 40.

51 *2019–2021 UN GGE Report*, n. 53, 10/26.

52 *Tallinn Manual 2.0* (n 50) Rule 6, 33.

53 ibid, Rule 6, 40.

54 See generally, Manos Antonakakis et al. (2017) 'Understanding the Mirai Botnet' in *26th USENIX Security Symposium (USENIX Security 17)*, 1094–1095, < https://www.usenix.org/system/files/conference/usenixsecurity17/sec17-antonakakis.pdf> accessed 12 November 2022.

55 See, Scott J Shackelford, Scott Russell and Andreas Kuehn, 'Unpacking the International Law on Cybersecurity Due Diligence: Lessons from the Public and Private Sectors' (2016) 17(1) *Chicago Journal of International Law* 20.

but legally plausible (based on intelligence inputs, past incidents, political relations with the affected state, etc.).[56] The standard of proof for a victim state to establish the territorial state's constructive knowledge, as Delerue notes, would be 'very high'.[57] Further, as this paper has highlighted in extensive detail, 'actual' or 'constructive' knowledge about malicious cyber activity on one element of a heterogeneous, multi-layered network cannot reasonably be tantamount to such knowledge of a targeted cyber operation. The lack of network visibility and control over various network components makes such determination difficult.

ATD does not offer a panacea to the problems highlighted above but will help bring the mitigation and prevention of transboundary digital harm operationally closer to expectations generated by the due diligence obligation in international law. First, ATD applications specifically allow for the detection of malicious cyber activity that causes transboundary harm. Second, and related, the adoption of ATD applications will help align legal and technical thresholds of 'actual' and 'constructive' knowledge of transboundary harmful activity. Third, ATD applications will operationally support an expansive cyber due diligence rule, the scope of which is not limited to halting harmful activity but also includes also a responsibility to prevent such activity. And fourth, ATD applications are useful not only in instances where harmful cyber operations originate in a state's territory but also where the operations transit through it.

Provided network security interface standards (such as I2NSF) are interoperable, ATD applications that run on them can detect unusual cyber activity in any virtualized component of the surveilled network, irrespective of the territory where such component is located. This will prove beneficial both to victim states and originator states. Victim states can pinpoint with greater precision the transboundary source of harmful cyber activity. Even if they cannot exercise control over the compromised infrastructure, they can promptly notify the originator state. ATD applications will not only guide originator states to malicious cyber activity and compromised infrastructure within their territory but could also indicate, through deep packet inspection,[58] the transboundary targets of such cyber operations. Given that they rely on ML models, ATD applications 'learn' and discern patterns from previous incidents involving transboundary harm and accordingly notify administrators when they detect similarly anomalous behaviour – at an early stage – on the surveilled network. In this manner, they ensure states have timely access to information necessitating their exercise of due diligence, and thus bring forward legal standards of 'actual' and 'constructive' knowledge to meet new technical realities. The predictive capability of ATD

56 Russell Buchan, 'Cyberspace, Non-State Actors and the Obligation to Prevent Transboundary Harm' (2016) 21(3) *Journal of Conflict and Security Law* 429, 441–442; *Tallinn Manual 2.0* (n 49) Rule 6, 41.

57 François Delerue, *Cyber Operations and International Law* (1st edn, Cambridge University Press 2020) 367.

58 A Carrega et al., 'Situational Awareness in Virtual Networks: The ASTRID Approach', in *2018 IEEE 7th International Conference on Cloud Networking (CloudNet)*, 3 https://ieeexplore.ieee.org/document/8549540> accessed 11 December 2021.

applications not only enhances the knowledge of states but could also help expand the scope of the cyber due diligence principle to include monitoring of networks and prevention of harmful transboundary activity. In other words, ATD applications make it technically feasible not only to monitor heterogeneous digital networks for anomalous behaviour but also to analyse in advance the precise nature of the threat and the harm it is likely to cause.

Finally, ATD applications could also help states detect malicious activity transiting through territorial networks.[59] Such detection is often quite challenging, given that traffic routing is usually automated and a factor of speed and availability of network resources, rather than of conscious choice to steer malicious activity through the servers of a particular country.[60] Knowledge of transiting traffic is then a determinant of superior intelligence and technical capacity as well as geopolitical attributes associated with a country's location. This perhaps explains why countries like the Netherlands and Singapore – major data transit points that both enthusiastically champion the norm on the 'public core of the internet'[61] and call on states to protect the availability and integrity of internet infrastructure that has transnational functionality – differ in their views on due diligence. The Netherlands claims a cyber due diligence rule already exists, whereas Singapore, which has to contend with the geopolitical volatility and cyber 'insecurity' of Southeast Asia,[62] has been more cautious, calling for greater clarity on the 'degree of knowledge' implicated by a due diligence rule.[63] ATD applications, which already interface with virtualization services agnostic of territorial location, would be well equipped to detect anomalous behaviour transiting through network infrastructure of those services.

59 For views of legal scholars on the responsibility of transit states, see *Tallinn Manual 2.0* (n 49), 33–34; Eric Talbot Jensen and Sean Watts, "Cyber Due Diligence," (2021) 73 *Oklahoma Law Review* 645, 696, 70; August Reinisch and Markus Beham, 'Mitigating Risks: Inter-State Due Diligence Obligations in Case of Harmful Cyber Incidents and Malicious Cyber Activity – Obligations of the Transit State' (2015) 58 *German Yearbook of International Law*, 101.

60 There may, of course, exist scenarios where a transit country's servers could be specifically targeted by state and non-state actors to thwart attribution.

61 Alexey Trepykhalin and Veni Markovski, 'Country Focus Report: The Netherlands and the "Public Core of the Internet"' (ICANN, 2021) <https://www.icann.org/en/system/files/files/ge-008-28may21-en.pdf> accessed 12 November 2021; 'Singapore's Written Comment on the Chair's Pre-Draft of the OEWG Report', <https://front.un-arm.org/wp-content/uploads/2020/04/singapore-written-comment-on-pre-draft-oewg-report.pdf> accessed 9 January 2022.

62 'Southeast Asia: Cyber Threat Landscape' (FireEye) <www.fireeye.com/offers/rpt-sea-threat-landscape.html> accessed 9 January 2022.

63 'Official Compendium of Voluntary National Contributions on the Subject of How International Law Applies to the Use of Information and Communications Technologies by States', UN Doc. A/76/136 (2021), 84/142 <https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf> accessed 3 March 2022.

5. CONCLUSION

With rapid advancements in computing, it appears probable that ATD across heterogeneous networks will soon be a reality. ATD applications can help realize and expand the due diligence obligation of states by facilitating the early detection and notification of transboundary digital harm. Not only would ATD applications offer states greater visibility over hardware and software elements of territorial networks, but they would also automate the reporting of potentially malicious activity. Critically, ATD applications will be capable of detecting new threats and attacks by learning from anomalous behaviour or signatures previously observed on known attack surfaces and vectors. By operationalizing the 'knowledge' component of cyber due diligence, ATD applications thus raise the standard of care owed by states for not only stopping but also preventing transboundary digital harm.

However, these applications too come with their own share of security concerns. Malicious attackers could corrupt the training data used by ATD algorithms, generating false positives or misleading results, which in turn depletes public trust in such technologies. The ATD framework reviewed in this paper, including the ASTRID project, also depends on APIs, whose security could be compromised with grave consequences for the integrity of the network as a whole. Finally, it is also possible that repressive or autocratic states may force the adoption of ATD applications with a view to engaging in deep inspection of private networks for surveillance, under the garb of security or performance of due diligence obligations.[64] These factors have to be carefully weighed against the widespread adoption of ATD. Additionally, such technologies may be inaccessible to developing countries for reasons of cost or export control restrictions, resulting in the uneven development and laggard adoption of cyber due diligence. Nonetheless, ATD responds to a pressing need to monitor diverse and multilayered networks. The deployment of intelligent algorithms to monitor and detect cyber security threats will not only optimize resources but also reduce response times for private and public entities. In the process, they may transform what it means for states to exercise diligence in the performance of their international obligations in cyberspace.

ACKNOWLEDGEMENTS

The title of the paper is inspired by that of the article 'The Responsibility to Inspect: Due Diligence in Cyberspace', which appeared on the website of the Observer Research Foundation in July 2016.

64 Delerue (n 57) 360–362.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

Exploring Changing Battlefields: Autonomous Weapons, Unintended Engagements and the Law of Armed Conflict

Tsvetelina J. van Benthem
University of Oxford
United Kingdom

Abstract: Battlefields are undergoing profound changes. Promises of increased precision, along with novel ways of verifying targets and ensuring accountability, accompany the introduction of new technologies. At the same time, many states, international and non-governmental organizations have voiced concerns over the increased levels and new types of uncertainty that emerging technologies may bring. This paper focuses on one particular technological development – autonomous weapons systems – and one particular risk associated with it – the possibility of unintended outcomes in the targeting of persons or objects. The submission centres on the law of armed conflict and more specifically on the relationship between unintended engagements and two relevant primary rules: the prohibition on making civilians the object of attack and the prohibition of indiscriminate attacks. Both prohibitions require reasonable decision-making. What is deemed reasonable depends on the availability of information on the status of targets and the modalities of attack. Given the importance of information, it is argued that a bolstered set of positive obligations to take steps (such as obligations to take precautions and to review new weapons) can act as information-generators, increasing the amount of knowledge available to parties to conflict in the planning and conduct of attacks and to other law appliers in considering the legality of particular engagements.

Keywords: *attacks against civilians, autonomous weapons, indiscriminate attacks, law of armed conflict, legal reviews, precautions*

1. INTRODUCTION

In the conduct of attacks, decision-making processes and weapons systems often fail. But it matters greatly why and how they fail, and what the consequences of such failures are. It has by now become clear that technical malfunctions of weapons and the misidentification of targets have led to enormous civilian suffering in armed conflict.[1] In such cases, civilian harm is not *intended*: the outcome of the attack does not match the intent of the deployer. According to some,[2] intent is a precondition for violating fundamental rules of the law of armed conflict (LOAC), such as the prohibition on attacking civilians.[3] Others say that a better interpretation of this particular prohibition suggests an objective approach.[4] What is clear is that there are uncertainties surrounding the existence and content of particular elements, objective and subjective, of key rules of the legal regime. Often triggered by particular events[5] or the delivery of important judgments,[6] clusters of discussions bring to the fore significant disagreements over the content of LOAC. Recent inter-governmental discussions on autonomous weapons systems show that these disagreements will likely become even more pronounced once the rules start being applied to new technologies that alter the ways in which parties to conflict interact with their weapons.[7] Hidden underneath these contestations is, ultimately, uncertainty regarding the contours of legal rules. This is significant, as a lack of clarity in the elements of rules directly impacts the constraining function of the law.

The substantive legal discussion of the piece follows two conceptual stages: identification and application. The main aim of this submission is to *identify* the elements of two key prohibitions of LOAC – the prohibition on making civilians the object of attack and the prohibition of indiscriminate attacks. Once their contours are identified, the rules will be *applied* to examples involving autonomous weapons systems.

1 Gregory McNeal, 'Targeted Killing and Accountability' (2014) 102 Georgetown Law Journal 681, 738 and footnote 296.

2 William Boothby, *The Law of Targeting* (OUP 2012).

3 Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts ('AP I'), 8 June 1977, art 51(2).

4 Lawrence Hill-Cawthorne developed an argument against the existence of subjective elements in the prohibition of directing attacks against civilians in Lawrence Hill-Cawthorne, 'Appealing the High Court's Judgment in the Public Law Challenge against UK Arms Export Licenses to Saudi Arabia' (EJIL: Talk!, 29 November 2018) <https://www.ejiltalk.org/appealing-the-high-courts-judgment-in-the-public-law-challenge-against-uk-arms-export-licenses-to-saudi-arabia/> accessed 14 April 2022.

5 For instance, one of the recent vigorous discussions in this area was provoked by the downing of a Ukrainian airliner by Iran at the start of 2020. See Marko Milanovic, 'Mistakes of Fact When Using Lethal Force in International Law: Part I' (EJIL: Talk!, 14 January 2020) <https://www.ejiltalk.org/mistakes-of-fact-when-using-lethal-force-in-international-law-part-i/> accessed 14 April 2022.

6 Hill-Cawthorne (n 4).

7 The divergent approaches of states can be seen by comparing, for instance, the U.S. Proposals on Aspects of the Normative and Operational Framework submitted to the Group of Governmental Experts on Lethal Autonomous Weapons Systems, CCW/GGE.1/2021/WP.3, 27 September 2021 ('2021 GGE US Working Paper'), and the Joint Working Paper Submitted by the Bolivarian Republic of Venezuela on behalf of the Non-Aligned Movement (NAM) and Other States Parties to the Convention on Certain Conventional Weapons (CCW), CCW/GGE.1/2021/WP.8, 8 December 2021.

The analysis will proceed in five steps. First, it will examine the changing landscape of battlefields and the problems – old and new – that pervade them. Then, it will describe the current state of militarized autonomous applications. The third step will involve identifying the content of the relevant LOAC rules and applying these rules to instances of unintended engagements carried through autonomous weapons systems. Fourth, a way of thinking about unpredictability and unintended outcomes will be proposed through an interaction of positive and negative obligations. The final section will summarize and conclude.

2. CHANGING BATTLEFIELDS

Some characteristics of battlefields appear to be immutable. A degree of unpredictability seems inherent in conflict, and targeting decisions are not, and cannot be expected to be, made in conditions of perfect certainty or knowledge.[8] Military operations take place in conditions of danger, pressure and uncertainty. Prior to the First World War, the brunt of these conditions was borne by those fighting on the battlefield, but this changed quickly with the increase of artillery range and the practice of aerial bombardments.[9] Today, especially with conflicts primarily taking place in urban areas,[10] civilians are under an increasing threat of harm. Battlefields are often portrayed as submerged under a fog of uncertainty. Writing at the beginning of the 19th century, von Clausewitz explained that '[t]he great uncertainty of all data in War is a peculiar difficulty, because all action must, to a certain extent, be planned in a mere twilight, which in addition not infrequently – like the effect of a fog or moonshine – gives to things exaggerated dimensions and an unnatural appearance'.[11] Unpredictability has been a pervasive feature of armed conflict for centuries. According to some states, the introduction of new technologies comes with a promise of making battlefields more predictable. With satellite imagery, surveillance drones and automated tools for the analysis of data, one hopes that the fog of war can and will be lifted through the allocation of adequate and sufficient resources.

Moving beyond the degree of uncertainty that has always accompanied conflict, we see new sources of uncertainty making their appearance. When a party to conflict deploys a weapons system that can independently select and engage targets, that party may introduce a new risk: the risk of not fully understanding the autonomous process of selection and engagement and of straining the link between intent and outcome. With traditional weapons, failures can be explained more easily. Weather conditions

8 Michael Schmitt and Michael Schauss, 'Uncertainty in the Law of Targeting: Towards a Cognitive Framework' (2019) 10 Harvard National Security Journal 148, 156.

9 Yves Sandoz, Christophe Swinarski and Bruno Zimmermann, Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949 (Martinus Nijhoff Publishers 1987), para. 1866 ('Commentary to AP I').

10 ICRC, War in Cities Casebook, <https://casebook.icrc.org/highlight/war-cities> accessed 14 April 2022.

11 Carl von Clausewitz, *On War* (Project Gutenberg, 1999) 70.

may impact their trajectories or cartridge chambers can get corroded, affecting performance.[12] But with emerging autonomous capabilities in target selection and engagement, a new risk factor may be the unpredictability of the ways in which these systems can fail.

Uncertainty is a poor conductor of the deployer's intent. This raises an important question: how does the law regulate mistakes about the status of targets and about the performance of the means we use to reach those targets? Commentators analysing LOAC often use the phrase 'things go wrong' to describe scenarios of errors. Dinstein, for instance, writes that 'many things can go wrong in the execution of attacks, and, as a result, civilians are frequently harmed by accident',[13] illustrating this proposition through the following examples:

(a) Faulty intelligence may lead to mistaken targeting, the victims of which are civilians/civilian objects.
(b) Human error, for example, misreading of data, miscalculation in navigation, etc, or missiles that go off course, artillery shells that go astray.
(c) Weapon systems suffering from technical malfunctions (due to no human error).
(d) Inclement weather that may deflect the trajectories of projectiles.[14]

A chapter in Boothby's book on the LOAC is entitled 'When Things Go Wrong'.[15] Abstracted and distanced from human decision-making, this phrase seems to suggest that something impersonal – 'things' – is at fault and is leading to civilian harm. Framing these instances as accidental is already conditioning our thinking regarding their positioning under (or rather – out of) the scope of legal rules. Intuitively, this framing suggests that, though they may be tragic and unfortunate, they are not wrongful.

The context of emerging technologies has already brought such questions to the fore. According to the September 2021 Working Paper on autonomous weapons systems submitted by the US, '[u]nintended harm to civilians and other persons protected by IHL [international humanitarian law] from accidents or equipment malfunctions, including those involving emerging technologies in the area of LAWS, is not a violation of IHL as such'.[16] Indeed, the *fact* of unintended civilian harm may not in *itself* trigger a violation of the relevant rules, but aspects of a faulty decision-making process may.[17] As will be demonstrated in Section 4 of this paper, two key LOAC

12 Bob Orkand and Lyman Duryea, *Misfire: The Tragic Failure of the M16 in Vietnam* (Globe Pequot Press 2019).
13 Yoram Dinstein, *The Conduct of Hostilities* (CUP 2016) 398.
14 ibid.
15 William Boothby, *The Law of Targeting* (OUP 2021), Ch 25.
16 2021 GGE US Working Paper (n 7) Section B, letter (g).
17 Note the use of 'as such' in the quote.

prohibitions proscribe careless decision-making in the choice of targets and means and methods of attack. Before turning to the specification of these LOAC rules, it is important to discuss the characteristics of autonomous weapons systems and the concerns that have been voiced over their deployment in combat.

3. BATTLEFIELD AUTONOMY

Just as aerial warfare revolutionized 20th-century battlefields, so will the use of autonomous machine-driven decision-making revolutionize those of the 21st century. Of particular note is the gradual distancing – temporal, geographical, causal and moral – between human input and the specific decisions to engage particular targets. Together with this distancing comes a wider surface for the emergence of uncertainties – uncertainties over concrete targets, over the decision-making process itself, and over external factors in their relation to machine performance. Because of these new uncertainties and their direct connection to targeting mistakes, an overview of autonomous weapons capabilities is particularly appropriate in the context of unintended engagements under LOAC.

Autonomous weapons systems bring promises of overcoming challenges, faults and frailties that permeate contemporary armed conflicts. Imperfect or insufficient information often leads to mistakes in the identification of targets, leaving many civilians dead or injured.[18] A common argument in support of autonomous weapons systems is that they will ensure a more discriminate conduct of hostilities[19] through capacities to home over areas for days, gathering more information on potential targets.[20] Anger, fear, frustration, and stress could drive conduct that imperils or harms civilians.[21] The lack of emotion in the operation of weapons stands in stark contrast to a pessimistic vision of the human element, the latter being construed as a battlefield risk of its own right. In addition to these perceived benefits, autonomous weapons could arguably ensure continued operation in communications-degraded environments,

18 For a recent example, see the explanation that came following the August 2021 strike in Kabul: Pentagon Press Secretary John F. Kirby and Air Force Lt. Gen. Sami D. Said Hold a Press Briefing (Transcript, 3 November 2021) <https://www.defense.gov/News/Transcripts/Transcript/Article/2832634/pentagon-press-secretary-john-f-kirby-and-air-force-lt-gen-sami-d-said-hold-a-p/> accessed 14 April 2022.

19 Anastasia Roberts and Adrian Venables, 'The Role of Artificial Intelligence in Kinetic Targeting from the Perspective of International Humanitarian Law', in T Jančárková, L Lindström, G Visky, P Zotz (eds), *Going Viral* (13th International Conference on Cyber Conflict, 2021) 53.

20 U.S. Proposals on Aspects of the Normative and Operational Framework submitted to the Group of Governmental Experts on Lethal Autonomous Weapons Systems, CCW/GGE.1/2021/WP.3, 27 September 2021, para 7.

21 Considerations for the report of the Group of Governmental Experts of the High Contracting Parties to the Convention on Certain Conventional Weapons on emerging technologies in the area of Lethal Autonomous Weapons Systems on the outcomes of the work undertaken in 2017–2021, submitted by the Russian Federation to the GGE on LAWS, 27 September 2021, CCW/GGE.1/2021/WP.1, para 7.

thereby reducing risks to the party's own troops – all factors that explain the billions spent on the research and development of autonomous weapons systems.[22]

Autonomy comes in all shapes and forms, and autonomous functions are best understood as standing along an autonomy spectrum. There are a number of existing military capabilities that are already on this spectrum, such as the HARPY missile and the Long Range Anti-Ship Missile (LRASM). HARPY is an all-weather day/night fire-and-forget missile with a loitering munition that 'detects, attacks and destroys enemy radar emitters, hitting them with high hit accuracy'.[23] The LRASM is an anti-ship missile with increased capabilities to conduct autonomous targeting. LRASM targeting relies on on-board targeting systems, meaning that prior intelligence would not be required.[24] Moving beyond these two particular missiles, it is anticipated that states will move to the development of systems that are capable of independently selecting *and* engaging targets.[25] Here, a conceptual and factual rupture between the intent of the party to conflict and the actual strike may occur in ways that make the selection of targets by the system unpredictable to those deploying it. How such a rupture may occur will likely become visible in the coming years: deployments of autonomous weapons systems have already been documented.[26]

Discussions on the definition of autonomous weapons have yielded little fruit, including at the Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE on LAWS), a group operating within the ambit of the Convention on Certain Conventional Weapons. Approaches to defining the capabilities of these weapons vary by actor and can be explained by the desire of a particular actor to develop such systems. The lack of a precise definition of the capabilities of the systems need not hamper a discussion on their regulation, however. In fact, delegations at the GGE have already spent years discussing the applicability of international law to such systems, going some way to specifying the content of relevant legal standards. The same is true of reports of international organizations[27] and academic work. The

22 See, for instance, Michael Klare, 'Pentagon Asks More for Autonomous Weapons' (Arms Control Association, April 2019) <https://www.armscontrol.org/act/2019-04/news/pentagon-asks-more-autonomous-weapons> accessed 14 April 2022.

23 HARPY: Autonomous Weapon for All Weather (IAI) <https://www.iai.co.il/p/harpy> accessed 14 April 2022.

24 Description at <https://www.naval-technology.com/projects/long-range-anti-ship-missile/> accessed 14 April 2022.

25 From ICRC Position on Autonomous Weapon Systems (March 2021). According to their definition, '[a] utonomous weapon systems select and apply force to targets without human intervention. After initial activation or launch by a person, an autonomous weapon system self-initiates or triggers a strike in response to information from the environment received through sensors and on the basis of a generalized "target profile".'

26 Final report of the Panel of Experts on Libya established pursuant to Security Council resolution 1973 (2011) S/2021/229, 8 March 2021, p 17.

27 A great example of work seeking to clarify applicable legal standards is the ICRC and SIPRI June 2020 report 'Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control' – Vincent Boulanin, Moa Peldán Carlsson, Netta Goussac and Neil Davison, 'Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control' (2020) SIPRI, available at: <https://www.sipri.org/publications/2020/other-publications/limits-autonomy-weapon-systems-identifying-practical-elements-human-control-0> accessed 14 April 2022.

relevance of these discussions cannot be overstated. Clarity around the rules of LOAC can both exert a deterrence pull for military capabilities that would fall short of legal requirements and facilitate accountability efforts following a breach of the law.

For now, what is clear is that there is consensus around the applicability of international law to autonomous systems.[28] This is consonant with the dictum of the International Court of Justice in the *Nuclear Weapons* Advisory Opinion, which affirmed that LOAC applies to 'all kinds of weapons, those of the past, those of the present and those of the future'.[29] But to agree on the applicability of LOAC is the easy step. Agreeing on *how* the law applies is a much more complex exercise. It is to this exercise that we turn in the next section.

4. IDENTIFYING TWO LOAC PROHIBITIONS

What form of decision-making is prohibited in the conduct of hostilities? There are many rules that constrain the ways in which conflict is waged by protecting different categories of persons and objects: civilians[30] and civilian objects,[31] medical units and personnel,[32] the natural environment,[33] installations containing dangerous forces,[34] among others. The prohibitions on making civilians the object of attack and of indiscriminate attacks are fundamental to the protective goals of the legal regime.

Before turning to a granular assessment of each rule,[35] it may be helpful to start with the conclusion on their scope. The prohibition on making civilians the object of attack proscribes the unreasonable construction of targets. That is, its regulatory effect is directed at the decision-making through which targets of attack are selected. Closely connected to the principle of distinction, this prohibition deals with the process of determining the status of a particular person, that is, whether they are a civilian or combatant and whether there is any doubt regarding their status, and the subsequent process of deciding how to act based on that determination. In that sense, it is a prohibition that looks at the relationship between the decision-making of a party to conflict and the status of their object of attack. In contrast, the prohibition of indiscriminate attacks regulates decision-making by reference to the modalities of an attack. These modalities can relate to the means and methods of the attack, but also to

28 Guiding Principle (a), Annex IV to Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1/2019/3, 25 September 2019.
29 Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, [1996] ICJ Rep 226, para 86.
30 AP I (n 3) arts 48, 50, 51.
31 ibid, art 52.
32 ibid, arts 12, 15.
33 ibid, art 55.
34 ibid, art 56.
35 This analysis is conducted on the basis of the rules of Additional Protocol I and follows the rules of treaty interpretation, as codified in the Vienna Convention on the Law of Treaties – VCLT, 1155 UNTS 331, arts 31–33.

the nature of the target identified (whether it is specific) and the controllability of the effects of the attack.[36] It is again about how an attack is planned and launched, but this time with regards to the way in which it is carried out.

A. Making Civilians the Object of Attack

According to Article 51(2) of Additional Protocol I, the civilian population, as well as individual civilians, shall not be the object of attack.[37] What does it mean to make civilians the object of attack? The starting point of the inquiry is the text of the provision; the terms are to be interpreted in accordance with their ordinary meaning.[38] The phrase 'shall not be the object of attack' seems, at first glance, capable of accommodating even the widest view of the scope of the prohibition, namely that civilians shall not, as a matter of *fact*, be engaged as an object of attack. No emphasis is being placed on any particular subjective state that the attacker may have vis-à-vis the status of those attacked. However, the better interpretation of this text is that it has a dual focus. First, it emphasizes the position of those attacked, of their state of *being* the object of an attack. Second, it implies the *construction of an object of attack* – a construction that occurs within the decision-making process of the attacker. At the very least, then, it suggests that the attacker intends to carry out an attack against persons or objects that it has constructed as its desired target, as opposed to these persons being incidentally harmed by the attack.[39]

Context is highly important in the interpretation of this rule. To begin with, where the drafters of Additional Protocol I sought to incorporate specific elements of intention, they did so explicitly. Comparing the two sentences of Article 51(2) enables us to discern the difference between prohibitions drafted to contain a specific purposive element and provisions drafted without such an element. To illustrate, Article 51(2), in its second sentence, contains a further prohibition: 'Acts or threats of violence *the primary purpose of which is to spread terror* among the civilian population are prohibited.'[40]

Another key piece to the puzzle is the grave breaches provision of Additional Protocol I. The Geneva Conventions and Additional Protocol I identify a small set of violations, called 'grave breaches', as serious enough to give rise to specific repressive obligations

36 AP I (n 3) art 51(4)(c).
37 ibid, art 51(2) first sentence.
38 VCLT (n 35) art 31(1).
39 Support for this reading can be found in the ICRC Commentary to art 51: '1938 The first sentence gives substance to the principle of general immunity formulated in the preceding paragraph by explicitly prohibiting attacks directed against the civilian population as such, as well as against individual civilians. By using the words "directed" and "as such" it emphasizes that the population must never be used as a target or as a tactical objective.'

for states.[41] These breaches constitute war crimes and should be prosecuted by High Contracting Parties to the Conventions and Protocol.[42] According to Article 85,

> In addition to the grave breaches defined in Article 11, the following acts shall be regarded as grave breaches of this Protocol, when committed wilfully, in violation of the relevant provisions of this Protocol, and causing death or serious injury to body or health:
> (a) making the civilian population or individual civilians the object of attack[43]

What this provision does, at first glance, is to take a number of LOAC prohibitions, including the one in Article 51(2), and add two further elements – wilfulness and causation of harm – to qualify the acts as grave breaches of the Protocol. Importantly for our purposes, the provision seems to superimpose a subjective element – wilfulness – on the prohibition on making civilians the object of attack. Wilfulness is defined in the Commentary as intention or recklessness.[44] In short, making the civilian population or individual civilians the object of attack, when committed wilfully and causing death or injury, is a grave breach under Article 85. Conversely, the prohibition in Article 51(2) would seem to require *something less than wilfulness* – that is, less than intention or recklessness. Something less than wilfulness, then, could be either a form of negligence or any attack that, as a matter of *fact*, makes civilians its object. As already discussed, the fact of causing harmful consequences to civilians does not seem to be what the prohibition is directed at. The better interpretation, taking into account the reasoning derived from the grave breaches provision, is that the prohibition on making civilians the object of attack proscribes the unreasonable construction of targets, a standard that encompasses subjective elements of intention, recklessness and negligence.

A final element of the contextual interpretation is the presumption of civilian status contained in Article 50(1) of Additional Protocol I, which reads: 'In case of doubt whether a person is a civilian, that person shall be considered to be a civilian.' It is significant that this provision has consistently been read into the prohibition of making civilians the object of attack in the jurisprudence of international tribunals. This was true for the International Criminal Tribunal for the Former Yugoslavia (ICTY) (in *Galić*)[45] and it is now true of the International Criminal Court (in *Katanga*[46] and *Ntaganda*).[47] Although these tribunals are tasked with the determination of individual

40 Emphasis added.
41 Dieter Fleck, *The Handbook of International Humanitarian Law* (OUP 2021) VII, 21.10, para 5.
42 ibid.
43 AP I (n 3) art 85(3).
44 ICRC Commentary, para 3474.
45 ICTY, *Prosecutor v Galić*, Trial Chamber Judgment, 5 December 2003, IT-98-29-T, para 42 et seq.
46 ICC, *Prosecutor v Katanga*, Trial Judgment, ICC-01/04-01/07-3436, 20 April 2015, para 797 et seq, referring to the wording of the prohibition in international humanitarian law.
47 ICC, *Prosecutor v Ntaganda*, Trial Chamber Judgment, ICC-01/04-02/06-2359, 8 July 2019, para 744 and footnote, as well as para 916 and following paragraphs.

criminal responsibility rather than state responsibility, their analyses on this point are of relevance here. This is because they first examined the content of the prohibition codified in Additional Protocol I before turning to the additional elements required by the legal framework regulating international crimes.

What, then, does this presumption of civilian status entail? The presumption concerns 'persons who have not committed hostile acts, but whose status seems doubtful because of the circumstances. They should be considered to be civilians until further information is available, and should therefore not be attacked.'[48] This presumption has a test of reasonableness, as persons 'shall not be made the object of attack when it is not reasonable to believe, in the circumstances of the person contemplating the attack, including the information available to the latter, that the potential target is a combatant'.[49] It is not the actual knowledge of risk of civilian status that would breach the prohibition; the analysis turns on whether a reasonable person with the information available at the time would have formed doubts regarding the status of the target.

The object and purpose of Additional Protocol I confirm this reading. If protecting civilians through an exercise of constant care is part of the object and purpose of the treaty, and if the foundational principle of distinction is operationalized through, inter alia, this specific prohibition, then a carelessness in making decisions on the status of targets must entail a violation of LOAC.

In the traditional context, the process of target selection and engagement is, to a substantial degree, in the hands of humans. How the target is initially identified, what materials decision-makers have recourse to, and how assumptions are questioned within the process are all traditional human endeavours to reach a conclusion on the status of a target. That paradigm is changing with the introduction of autonomy on the battlefield, where target identification and selection decisions are increasingly being relegated to machine sensors and algorithms. Consider the 16 March 2002 bombing of the Donetsk Regional Theatre of Drama in Mariupol by Russian forces.[50] Civilians had been sheltering in the theatre for days and had inscribed 'ДЕТИ' – Russian for 'children' – on both sides of the building in an attempt to ensure that attacking forces would abstain from making it an object of attack. At the very least, this inscription ought to have created doubt in the mind of the attackers, thereby requiring them to abstain from its engagement unless conclusive information was gathered to the effect that it was, in fact, a military objective and its attack was not expected to cause harm excessive to the anticipated military advantage. For human operators, it is clear that a sign 'ДЕТИ' would communicate information that ought to shape their decision-

48 ICRC Commentary, para 1920.
49 *Galić* (n 45) para 50.
50 Hugo Bachega and Orysia Khimiak, 'Mariupol theatre: "We knew something terrible would happen"' (BBC News, 18 March 2022) <https://www.bbc.co.uk/news/world-europe-60776929> accessed 14 April 2022.

making, but would this be true for an autonomous weapons system? How would it perceive such a sign – would it perceive it at all? As these weapons rely on sensors to identify persons and objects, on processing systems to compartmentalize what has been observed through the sensors and on algorithms to direct the initiation of attack upon the fulfilment of necessary conditions,[51] the protection of civilians would rely on the individual performance of each of those systems, their rigorous and comprehensive programming and smooth interaction. Translating human cognition to machine situational awareness will be a complex task, and one that will be accompanied by substantial risks for civilians.

When a commander or operator is deploying an autonomous system that can independently select and engage targets, the *way* in which that system selects its objectives is part and parcel of the method through which the party to conflict constructs its targets. In practical terms, parties to conflict should ensure that the system is equipped with sufficient sensors and analytical nodes to determine its targets with the required degree of certainty, that the presumption of doubt is translated into the algorithm of the weapon, and that biases in the collection and analysis of information are eliminated or mitigated. Just as the members of the target cell in the August 2021 Kabul strike suffered from confirmation bias,[52] algorithms can become biased through the biased information that they are trained on.[53] If a weapon's algorithms suffer from bias and those algorithms are used to channel the intent of the party to conflict, then it cannot be said that the party is complying with its obligation to not make civilians the object of attack. While much of the groundwork on ensuring the safety and predictability of autonomous systems will occur at a much earlier stage (research, design, development), the metric by which deployment is judged is similar to that in the traditional context, and a party to conflict can violate the prohibition on attacking civilians when the system is unable to operate with a sufficient basis of certainty in target identification and verification.

States have already proffered views on how the safe deployment of autonomous systems could be ensured. The US, for instance, has emphasized the role of target verification and evaluation, testing and training, understandable weapon interface, and training on system activation and deactivation.[54] A Joint Working Paper submitted to the GGE on LAWS by the Argentine Republic, the Republic of Costa Rica, the Republic of Ecuador, the Republic of El Salvador, the Republic of Panama, the State of Palestine, the Republic of Peru, the Republic of the Philippines, the Republic of

51 MT Klare, 'Autonomous Weapons Systems and the Laws of War' (Arms Control Association, 2019) <https://www.armscontrol.org/act/2019-03/features/autonomous-weapons-systems-laws-war> accessed 14 April 2022.

52 Press briefing (n 18).

53 Twitter already discovered the importance of data fed into its algorithms: 'Twitter's racist algorithm is also ageist, ableist and Islamaphobic, researchers find' (NBC News, 9 April 2021) <https://www.nbcnews.com/tech/tech-news/twitters-racist-algorithm-also-ageist-ableist-islamaphobic-researchers-rcna1632> accessed 14 April 2022.

54 2021 GGE US Working Paper (n 7) para. 10.

Sierra Leone and the Eastern Republic of Uruguay emphasized transparency, review procedures and public and industry safeguards.[55] States are gradually moving towards the specification of an operational framework.

B. The Prohibition of Indiscriminate Attacks

A wide pool of wrongful behaviour is covered by the prohibition of indiscriminate attacks,[56] which seeks to constrain certain modalities of attack that contravene the principle of distinction. Indiscriminate attacks are related to certain characteristics of the target chosen (its specificity) and the means and methods used in the attack, including whether the effects of the attack can be limited. It is the employment of indiscriminate tactics or weapons that is prohibited under this heading.[57] What indiscriminate attacks betray is a disregard for civilian life by the decision-maker.[58] Examples of indiscriminate attacks include the random discharge of bombs, careless firing without checking one's targets, targeting in conditions of impaired visibility, be it due to high altitude or inclement weather conditions, and the employment of imprecise weapons in densely populated areas.

Article 51(4)(b) prohibits a particular type of disregard for civilian life that is of utmost relevance to the present analysis: the use of certain means of warfare that 'are of a nature to strike military objectives and civilians or civilian objects without distinction'.[59] Both weapons that are intrinsically indiscriminate and those that are inaccurate (though not indiscriminate per se) have been examined under this prohibition.[60] To begin with, certain weapons are considered per se indiscriminate as they cannot be aimed at specific targets. A paradigmatic example is the German V2 rocket used during the Second World War. Even if a weapon is not considered indiscriminate in all circumstances, it can be deemed indiscriminate for a particular context. In the *Martić* trial in front of the ICTY, the Trial Chamber found that the use of non-guided rockets with cluster munitions aimed at the city of Zagreb from 50 kilometres away constituted an indiscriminate attack.[61] The *Gotovina* Trial and Appeals Judgments of the ICTY highlighted the difficulties of establishing clear standards for the margin of error of weapons, in that case, of artillery weaponry. The ICTY Trial Chamber attempted to draw a line in the sand by introducing a 200-metre test for assessing the intended targets of artillery projectiles. More precisely, it

55 2021 Joint Working Paper submitted by the Argentine Republic, the Republic of Costa Rica, the Republic of Ecuador, the Republic of El Salvador, the Republic of Panama, the State of Palestine, the Republic of Peru, the Republic of the Philippines, the Republic of Sierra Leone and the Eastern Republic of Uruguay to the GGE on LAWS, CCW/GGE.1/2021/WP.7, 27 September 2021 ('GGE Working Paper 7').

56 AP I (n 3) art 51.

57 MN Schmitt, 'International Humanitarian Law and the Conduct of Hostilities' in B Saul and D Akande (eds), *The Oxford Guide to International Humanitarian Law* (OUP 2020) 152.

58 H.M Hanke, 'The 1923 Hague Rules of Air Warfare' (1993) 33 IRRC 12, 26.

59 AP I (n 3) art 51(4).

60 C Ponti, 'The Crime of Indiscriminate Attack and Unlawful Conventional Weapons: The Legacy of the ICTY Jurisprudence' (2015) 6 Journal of International Humanitarian Legal Studies, 118, at 136. Though note that the ICTY did not adjudicate on indiscriminate attacks as an autonomous offence.

61 ICTY, *Prosecutor v Martić*, Trial Chamber Judgment, IT-95-11-T, 12 June 2007, paras 462–463.

considered that artillery attacks impacting anything within 200 meters of an identified artillery target had been fired at that target. Artillery attacks outside the radius were, in contrast, viewed as evidence of an indiscriminate attack. The Appeals Chamber rejected that test, finding no basis for it in law, and emphasized a case-by-case analysis moulded to the particular circumstances of a strike.[62]

There is no exhaustive list of weapons considered inherently indiscriminate, nor is there a list of weapons the use of which is automatically considered indiscriminate in particular battlefield contexts. Nevertheless, benchmarks are gradually forming through the engagement of states, tribunals, international organizations and academia. Three benchmarks are of particular relevance to the application of this prohibition to emerging technologies.

First, despite the lack of precise standards in treaty law or customary international law, certain trends are starting to surface. For instance, in the 2019 Report of the Group of Eminent International and Regional Experts on Yemen, we read that 'certain types of explosive weapons with a wide impact area, such as artillery, mortars, and unguided rockets, which use blast and fragmentation to kill and injure are inherently inaccurate when used in populated areas'[63] – a statement for which the Group derives support from a prior International Committee of the Red Cross (ICRC) report. It may be that, over time, uses of certain weapons in particular contexts will become 'givens' of illegality.

Second, standards of what is unlawful evolve: '[T]echnological developments ... may shift general understandings as to when a weapon is incapable of being directed. In particular, improvements in the accuracy of weapons may heighten expectations of the general public as to precision.'[64]

Third, while the wording of the prohibition of indiscriminate attacks and its application to date do not seem to require proof of any subjective elements of cognition regarding the technical performance of weapons systems, this may be due to an underlying assumption that parties to conflict have a good understanding of how their means of warfare operate. This assumption may be challenged in the future with the introduction of increasing levels of autonomy in machine decision-making. Though it is here proposed that an objective standard for determining whether the employment of a particular weapon was reasonable in the circumstances better aligns with both the textual and the teleological interpretation of the rule enshrined in Article 51(4), that reasonableness itself will necessarily turn on information available on the particular type of weapons system, its functioning, interactions and effects.

62 ICTY, *Prosecutor v Gotovina*, Appeals Judgment, IT-06-90-A, para 60.

63 Report of the detailed findings of the Group of Eminent International and Regional Experts on Yemen, 3 September 2019, A/HRC/42/CRP.1, para 316.

64 Program on Humanitarian Policy and Conflict Research at Harvard University, Commentary on the HPCR Manual on International Law Applicable to Air and Missile Warfare 64 (CUP 2010).

For now, it seems clear that not every time a weapon system misperforms or weather conditions alter the trajectory of missiles will we be able to say that a party initiated an indiscriminate attack. As argued by Simpson and Müller, all weapons have tolerance-level specifications that determine the accepted degree of reliability they must possess.[65] Specifying these accepted degrees of reliability for different deployment scenarios will be key. Complex questions around the accuracy and reliability of weapons are likely to emerge with the introduction of autonomous weapons systems on the battlefield. A particular concern in the area of autonomous weapons is their feared propensity for unpredictable and unexplainable operation. Echoing the recently published position of the ICRC,[66] a 2021 Joint Working Paper submitted to the GGE on LAWS opines that,

> Unpredictable autonomous weapon systems should be expressly ruled out, notably because of their indiscriminate effects. In other words, a prohibition on autonomous weapon systems that are designed or used in a manner such that their effects cannot be sufficiently understood, predicted and explained should be prohibited under international law.[67]

For now, there is a dangerous pattern: the more independent machines are, and the more complex their machine learning algorithms, the more opaque they are to us.[68] It is entirely possible to envisage a future where these systems will become, through repeated testing and training, predictable to their deployers. This does not, however, mean that they will not have the capacity to fail just as any other weapon does. When a malfunction is observed, the inquiry into the potentially indiscriminate nature of the deployment will depend on the context, the weapon's specifications, its performance in previous deployments, as well as adequacy of data training for the particular battlefield environment.

5. POSITIVE AND NEGATIVE OBLIGATIONS, AND THE IMPORTANCE OF INFORMATION

Amid the chaos of conflict, information is one of the most precious resources. Information constructs what a reasonable party would have done in particular circumstances. The availability of information on the performance of weapons, the processes through which they operate, the ways to exercise control over their function, and the method for selecting targets and operationalizing the presumption of civilian status are crucial pieces of the analysis. Decision-makers *should know*

65 Thomas W Simpson, Vincent C Müller, 'Just War and Robots' Killings' (2016) 66 The Philosophical Quarterly 263, s 4.

66 See n 24.

67 GGE Working Paper 7 (n 52).

68 Kartik Hosanagar, 'As machines become more intelligent, they also become unpredictable' (2 August 2019) <https://www.foundingfuel.com/article/as-machines-become-more-intelligent-they-also-become-unpredictable/> accessed 14 April 2022.

more, as knowledge of what they deploy and how their systems operate is critical to ensuring civilian protection. How can parties to conflict be made to know more? LOAC contains a range of obligations that require them to take affirmative steps, such as the precautionary obligations of target verification and care in the choice of means and methods of attack,[69] as well as the obligation to review new weapons.[70] In the process of complying with these obligations, parties to conflict produce new information on their weapons, their expected performance and the target verification processes. All this data then becomes part of the pool of available information. And this is how the interaction between positive and negative obligations occurs: positive obligations, in generating information, impact what is expected of parties under the two prohibitions examined in this article. States should thus not lose sight of the importance of the obligations to take positive steps. They have a crucial role to play in operationalizing protection under core LOAC negative duties.

6. CONCLUSION

Conflict is riddled with factual uncertainty, and this uncertainty places the life and limb of civilians at risk. But there is a different type of uncertainty that adds yet another dimension of risk: legal uncertainty, that is, uncertainty around the scope of rules from the LOAC framework. Emerging technologies may further strain the regime, as they raise issues in the grey areas of rules – areas that have never been sufficiently specified. The aim of this article was to clarify one set of legal uncertainties – that of the regulation of unintended engagements under two prohibitions of LOAC, the prohibitions on making civilians the object of attack and indiscriminate attacks – in their traditional form and in their application to emerging autonomous capabilities.

Deployments that properly conduct intent are in everyone's interest. In the coming years, states and other stakeholders should focus their efforts on the processes that can ensure safe, informed and understandable interactions between decision-makers and their weapons. The law, properly interpreted and applied, can – and should – play a key role in regulating these interactions.

69 AP I (n 3) art 57(2)(a)(i)–(ii).
70 ibid, art 36. It is important to note that the customary character of this AP I prohibition is contested.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

Legal Aspects of Misattribution Caused by Cyber Deception

Petr Stejskal
Palacky University Olomouc,
Faculty of Law, Centre for
International Humanitarian
and Operational Law

Martin Faix
Palacky University Olomouc,
Faculty of Law, Centre for
International Humanitarian
and Operational Law

Abstract: This contribution introduces the concept of cyber misattribution caused by deception. As cyber threats and tactics of their originators develop, so must international law keep moving and prevent exploitation of the rule of law by removing legal gaps surrounding deceptive actions of States. Cyber deception refers to a situation when a State launches a false-flag cyber attack against another State but orchestrates the attack in a way that points towards a third (victim) country as the wrongdoer. The target State then launches retaliatory measures against the alleged wrongdoer. The legal analysis of the proposed contribution focuses on the legality of such deception and responsibility of both its author and the deceived State for damage caused to the victim State. The contribution demonstrates the gap in the rules of international responsibility for holding the orchestrator of deception responsible for the damage caused to a victim State as a consequence of misattribution. It also focuses on the legality of the deception as such. Misleading another State is a matter not per se regulated by international law, but it may result in a violation of the no-harm principle. This principle is recognized as a distinct legal norm in specific areas, but it is unclear whether and how it applies to the cyber domain. Finally, the contribution analyses whether the responsibility of the deceived State may be alleviated based on a mistake of fact that caused the misattribution.

Keywords: *false-flag attack, deception, misattribution, mistake of fact, counter-measures, derived international responsibility*

1. INTRODUCTION

Deception is human activity that can take place in various forms or contexts whenever humans interact. This article introduces the concept of cyber deception aimed at causing misattribution and damage to a victim State as a more advanced (in respect of a plurality of parties as well as technical sophistication) form of deception. Apart from the law of armed conflict, deception on the inter-state level is not regulated by international law and does not enjoy the attention of the doctrine of international law so far. However, deception in the cybersphere is a threat and sophisticated challenge for international security due to the technical specifics of this domain and the opportunities this tactic offers. Resorting to false-flag attacks is a matter of concern also for intelligence services[1] and the North Atlantic Treaty Organization (NATO) is committed to addressing modern cyber threats in general.[2] NATO CCDCOE also prepared a report on mitigation of risks arising from false-flag and no-flag cyber attacks.[3] This paper, to the best of the knowledge of the authors, can be considered as one of the first academic legal works that address this phenomenon from the perspective of public international law. It aims to present and legally assess hypothetical deceptive cyber operations intended to damage another State and to initiate expert discussion on this emerging threat.

Cyber deception, as introduced in this contribution, refers to the tactics when a deceiving State launches a cyber operation against another State but orchestrates the attack in a way that points towards a third (victim) country as the wrongdoer. The deceived State may eventually launch retaliatory measures against the alleged wrongdoer, who naturally denies any responsibility for the cyber attack. This can lead to destabilization and conflict between the deceived and the victim States at potentially little cost (political, material and legal) to the deceiving State, unless it is revealed. Confusion and exploitation of the so-called attribution problem may also be examples of the lawfare tactics in the broader hybrid warfare strategy.[4]

The following analysis focuses on the legality of the deceptive conduct orchestrated by the deceiving State with the intent of causing damage to the victim State. As part of that, rules for invocation of responsibility for the conduct of another State

1 See, for example, the 2020 annual report of the Security Information Service (Czech intelligence and counterintelligence service), 'Annual Report of the Security Information Service for 2020' (2021) 18. The US National Security Agency (NSA) and UK National Cyber Security Centre (NCSC) even established a joint advisory group to avoid misattribution. See 'NSA and NCSC Release Joint Advisory on Turla Group Activity' (CISA, 23 October 2019) <https://www.cisa.gov/uscert/ncas/current-activity/2019/10/21/nsa-and-ncsc-release-joint-advisory-turla-group-activity> accessed 12 January 2022.

2 NATO, 'Wales Summit Declaration' (2014) 73.

3 Mauno Pihelgas, 'Mitigating Risks arising from False-Flag and No-Flag Cyber Attacks' (2015) <https://ccdcoe.org/library/publications/mitigating-risks-arising-from-false-flag-and-no-flag-cyber-attacks/> accessed 12 January 2022.

4 Tomáš Bruner and Martin Faix, 'The Attribution Problem as a Tool of Lawfare' (2018) 18(1) Obrana a strategie 79.

are examined. Then, the analysis addresses legal responsibility for cyber retaliation directed against the wrong target due to misattribution. Whether a mistake of fact can alleviate the international responsibility of the deceived State towards the victim State is examined. This paper does not focus in detail on the rules of attribution.[5] The following analysis demonstrates that rules for international responsibility are not sufficient to invoke responsibility on a State that misled another State and caused damage to a victim State. The paper also shows that an error on the side of the misled State does not generally alleviate its responsibility for the retaliatory measures against the alleged wrongdoer. The paper presents the view that the act of deception as such is not in contradiction with any of the established primary rules of general international law.

Based on that, the contribution calls for further academic discussion and points to other legal concepts that might apply to the act of deception. As cyber threats and the tactics of their originators develop, so must international law keep moving and prevent exploitation of the rule of law by removing legal gaps surrounding deceptive actions.

2. CONCEPT OF CYBER DECEPTION

Deceptions are commonly encountered in everyday life in economic interactions, sport, politics, diplomacy. Deception can be defined as an interaction between two parties in which the deceiver successfully causes the target to accept as true a specific incorrect version of reality, with the intent of causing the target to act in a way that benefits the deceiver.[6]

On various levels of inter-state relations, deception may also be utilized as a strategy by States pursuing their interests (be it on the level of diplomacy, negotiation, the conclusion of strategic agreements, etc.). Naturally, deception is also a method used for gaining a military advantage in the theatre of operations during hostilities. From a different angle, the concept of deception is also an important part of the cyber security defence strategies in the information technology sector.[7] This article introduces and focuses on the concept of cyber deception aimed at causing misattribution and damage to a victim State as a more advanced (in respect of a plurality of parties as well as technical sophistication) form of deception.

5 The issue of attribution is even more challenging in situations where a State uses non-state actors as proxies for the conduct of malicious cyber operations.

6 Neil C Rowe and John E Custy, 'Deception in Cyber-Attacks' in Andrew Colarik and Lech Janczewski (eds), *Cyber War and Cyber Terrorism* (Idea Group 2007) 91.

7 Kristin E Heckman et al. (eds), *Cyber Denial, Deception and Counter Deception: A Framework for Supporting Active Cyber Defence* (Springer 2015) – addressing deception as a strategy and technical method of resisting and eliminating cyber intrusions.

In a hypothetical scenario, a deceiving State may decide to conduct a harmful cyber operation against another State and include deceptive elements to orchestrate wrongful attribution of this operation to a third, victim State. The aim of the deceiving State would be that misattribution of such a false-flag cyber operation would be followed by retaliatory measures of the deceived State against the victim State. If the desired consequences of the false-flag operation materialize, the deceived State can resort to various measures against the alleged wrongdoer, ranging from diplomatic and political actions, economic countermeasures, to retaliatory measures in cyberspace. This can lead to destabilization and conflict between the deceived and the victim States with potentially little cost (political, material and legal) for the deceiving State, unless it is revealed.

There are multiple technical methods of execution of this *modus operandi*. With respect to the target of the false-flag cyber operation, examples from the practice show that the malware used against the deceived State may be designed to cause significant damage, for example, to the State's critical infrastructures. An example of such damage comes from 2015 when Ukraine's electric power grid was hit by damage to servers and workstations of a national control centre and sub-centres. The preparation phase of the attack was devoted to the spear-phishing campaign, infecting relevant machines and opening a backdoor for the destructive malware by deleting critical system files, which finally resulted in temporary loss of control over the distribution of power across the country and local power outages.[8] Targeting of infrastructures, such as the electric power grid, is especially perilous, as severe damage may lead to a chain of further consequences, such as the inoperability of essential State infrastructure or government services.

Concerning the methods used for deception, the cyberspace domain offers various technical possibilities for concealment and deception.[9] For this primarily legal study, it is sufficient to mention that the practice has already shown that malware used for false-flag cyber arrack may be created in such a way as to point towards another State or actor. The deceiving State may include specific segments of codes and combine them in a way to allow the deceived State to identify false, but persuasive, traces leading to the victim State.[10] The alleged origin of the cyber incident may also be indirectly supported by false information acquired by intelligence agencies and by the contextual elements (e.g. if the relations between the deceived and the victim States are unfriendly in general or if there has already been a real incident between them)

8 See for example Kim Zetter, 'Inside the Cunning, Unprecedented Hack of Ukraine's Power Grid' (*Wired*, 3 March 2016) <https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/> accessed 1 September 2021.

9 See for example Florian Skopik and Timea Pahi, 'Under false flag: using technical artifacts for cyber attack attribution' (2020) 3 Cybersecurity <https://cybersecurity.springeropen.com/articles/10.1186/s42400-020-00048-4> accessed 10 January 2022.

10 ibid 14.

as attribution in cyberspace is mainly based on technical analysis of the attack and all-source analysis.[11]

Due to its design (where deception is successful) or because of the policy and security considerations of the targeted State that revealed the true origin of the false-flag attack, the frequency of incidents of false-flag attacks is hidden under the veil of mystery. However, as mentioned previously, the practice has already shown some relevant examples of cyber deception utilised by State or State-related actors that were made public. There are reported incidents where China or Russia designed their false-flag attack in such a way as to point to Iran[12] or a non-State actor.[13] Probably more publicly known is the Olympic Destroyer malware, which targeted computers used by officials, athletes and visitors during the 2018 Winter Olympic Games in Pyeongchang.[14] The operation bore several features previously attributed to cyber espionage and sabotage actors allegedly based in or working for China and North Korea (known as the Lazarus group).[15] Subsequently, it was revealed that a Russian foreign military intelligence (GRU) unit was behind the operation.[16]

A successful deception operation and the materialization of its purpose consists of two distinct acts of two actors: the false-flag operation carried out by the deceiving State against the deceived State (incident 1) and possible retaliatory measures of the latter against the victim State as the alleged wrongdoer (incident 2). The following text will legally analyse both elements.

11 The 2020 annual report of the Security Information Service (n 1) 18. In fact, identified actions are correlated with socio-political contextual indicators and known capabilities and tactics of actor groups, meaning that attribution is often based on matching the investigation findings with threat actor profiles. See Skopik and Pahi (n 9) 4, 10–11.

12 Pierluigi Paganini, 'UK/US investigation revealed that Russian Turla APT masqueraded as Iranian hackers' (*securityaffairs.co*, 21 October 2019) <https://securityaffairs.co/wordpress/92770/apt/turla-false-flag-iran.html> accessed 12 December 2021.

13 Raphael Satter, 'Russian hackers posed as IS to threaten military wives' (*AP News*, 8 May 2018) <https://apnews.com/article/mi-state-wire-or-state-wire-russia-co-state-wire-north-america-4d174e45ef5843a0ba82e804f080988f> accessed 12 December 2021. Also interesting is the TV5Hack, initially pointing to the Cyber Caliphate group, but later assigned to the Russian APT28 group. See Skopik and Pahi (n 9) 14–17.

14 U.S. Department of Justice, 'Six Russian GRU Officers Charged in Connection with Worldwide Deployment of Destructive Malware and Other Disruptive Actions in Cyberspace' (*justice.gov*, 19 October 2020) <https://www.justice.gov/opa/pr/six-russian-gru-officers-charged-connection-worldwide-deployment-destructive-malware-and> accessed 9 December 2021.

15 'The Olympic False Flag: How infamous OlympicDestroyer malware was designed to confuse cybersecurity community' (*Kaspersky*, 8 March 2018) <https://www.kaspersky.com/about/press-releases/2018_the-olympic-false-flag> accessed 9 December 2021.

16 ibid; National Cyber Security Centre, 'UK and partners condemn GRU cyber attacks against Olympic and Paralympic Games' (*ncsc.gov.uk*, 19 October 2020) <https://www.ncsc.gov.uk/news/uk-and-partners-condemn-gru-cyber-attacks-against-olympic-an-paralympic-games> accessed 9 December 2021.

3. LEGAL EFFECTS OF CYBER DECEPTION (INTERNATIONAL RESPONSIBILITY FOR THE CONDUCT OF THE DECEIVED STATE)

Under current international law, it is only the law of armed conflict that addresses specific forms of deception. Article 37 of the Additional Protocol I to the Geneva Conventions and Article 21 of the Additional Protocol II to the Geneva Conventions state that ruses of war are not generally prohibited. However, these provisions and customary international law prohibit perfidy.[17]

In general, the question of legal consequences of a false-flag cyber operation must be divided into two separate questions: responsibility for the conduct against the deceived State (depending on the nature of the false-flag attack) and responsibility for the damage caused to the victim State.

Concerning the former in cases of an attack against a power grid, launching a cyber operation that causes loss of control over the distribution of power across the territory of another State and local power outages can lead to the loss of functionality of that State's critical infrastructure and e-government systems, significantly reducing its capability to serve its inherently governmental functions and its ability to conduct its affairs freely. Such a cyber operation may therefore amount to a breach of the prohibition on intervention and the obligation to respect the sovereignty of other States.[18] According to some experts, the mere disruption and malfunctioning of the cyberinfrastructure of another State may already qualify as a violation of territorial sovereignty.[19] The fact that causation of physical consequences even by remote means can be inconsistent with these rules seems to be generally accepted.[20] Moreover, if an armed conflict exists between the two States, an extensive attack on the electric power grid may also result, depending on circumstance, in a violation of the law of armed conflict. However, a more challenging question arises as to whether the State orchestrating the deception may be internationally responsible for the wrongful conduct committed by the deceived State towards the victim State. In other words, is it possible, and if so, on what grounds, for the victim State to invoke the responsibility of the author of the deception for the consequences caused by the retaliatory measures of the deceived State? The rules on State responsibility contain legal constructions

17 'Practice Relating to Rule 65. Perfidy' (ICRC) <https://ihl-databases.icrc.org/customary-ihl/eng/docindex/v2_rul_rule65> accessed 9 January 2022.

18 Michael N Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (CUP 2017), commentary to Rule 4 at para 13. See also the *Case Concerning Military and Paramilitary Activities in and against Nicaragua (Nicaragua v United States of America)*, International Court of Justice, Judgment (1986) at para 205 and *Tallinn Manual 2.0* commentary to Rule 66 at paras 15–18.

19 *Tallinn Manual 2.0* (n 18) commentary to Rule 4, para 13.

20 ibid, para 11.

21 *Tallinn Manual 2.0* (n 18) Rule 18.

on how a State can incur responsibility in connection with the wrongful conduct of another State. These rules also apply to cyberspace.[21]

According to the principle of independent responsibility, each State is responsible for its own internationally wrongful conduct.[22] However, a joint or collective wrongful act may lead to a plurality of responsible States.[23] Moreover, a State may also be responsible for a wrongful act of another State if it is implicated in the conduct of that other State. International law recognizes several forms of derived international responsibility:[24] aid or assistance (to assist in the commission of a wrongful act by another State),[25] direction or control (over the commission of an internationally wrongful act of another State)[26] and coercion (of another State to commit an internationally wrongful act).[27] These forms of implication have in common that the specific nature of the relationship between the State that is the actual author of the unlawful act and the implicated State causes the incurrence of responsibility of the latter.[28] The assisting State will typically not be responsible for the assisted wrongful act,[29] but for a distinct wrongful act – that is, deliberately assisting another State in breaching an international obligation by which they are both bound.[30] In contrast, the exercise of direction and control or coercion by one State over the commission of an internationally wrongful act by another incurs responsibility for the (directed or coerced) act itself[31] towards the injured State.[32] The coerced State might benefit from *force majeure* if the requirements are met.[33] In that case, it would be only the State exerting coercion that would bear responsibility.[34]

22 James Crawford, *State Responsibility: The General Part* (CUP 2013) 333; Yearbook of the International Law Commission, 2001, vol. II, Part Two - *Draft Articles on Responsibility of States for Internationally Wrongful Acts*, with commentaries. Commentary to Part IV at para 1.

23 See Christian Dominicé. 'Attribution of Conduct to Multiple States and the Implication of a State in the Act of Another State' in James Crawford et al. (eds), *The Law of International Responsibility* (OUP 2010) 282–284.

24 Crawford (n 22) 336.

25 Draft Articles (n 22) art 16. This concept was applied by the ICJ in the *Bosnian Genocide Case, see Case Concerning Application of the Convention on the Prevention and Punishment of the Crime of Genocide* (Bosnia and Herzegovina v Serbia and Montenegro), ICJ, Judgment (2007) para 420.

26 Draft Articles (n 22) art 17. This form of indirect responsibility is rare, belligerent occupation being one of the few examples. A distinction must be made from the situation where an organ of one State has been placed at the disposal of another State. Upon certain conditions, acts of this organ might be attributable to the latter State. See *Tallinn Manual 2.0* (n 18) Rule 16.

27 Draft Articles (n 22) art 18.

28 Christian Dominicé, 'Attribution of Conduct to Multiple States and the Implication of a State in the Act of Another State' in James Crawford et al. (eds), *The Law of International Responsibility* (OUP 2010) 284.

29 In situations where aid or assistance is an essential and integral element of the assisted State's operation, the assisting State may be responsible for the assisted conduct. Responsibility of the assisting State therefore attaches for the extent of its contribution. See *Tallinn Manual 2.0* (n 18) commentary to Rule 18, para 6.

30 Dominicé (n 28) 285; Draft Articles (n 22) commentary to art 16, para 10.

31 Draft Articles (n 22) commentary to art 17, para 1, commentary to art 18, paras 1 and 7; *Tallinn Manual 2.0* (n 18) commentary to Rule 18, para 6.

32 Dominicé (n 28) 288.

33 Draft Articles (n 22) commentary to art 23, para 3.

34 Dominicé (n 28) 288-289.

The problem is that the commission of the deceptive false-flag cyber operation as described above does not fit any of the recognized forms of the implication of international responsibility for the conduct of the deceived State. The nature of the relationship between the deceiving and the deceived States does not qualify as aid or assistance since the deceived State was not aware of the origin of the false-flag operation and intent of the deceiving State. In the situation where the deceived State misattributes the false-flag attack, there is also probably no relationship of dependence that would amount to direction or control.[35] Finally, that State is also not typically coerced to engage in retaliatory measures against the alleged wrongdoer, as it was not deprived of its freedom of action.[36] Consequently, the victim State cannot invoke the responsibility of the deceiving State for the consequences caused by the retaliatory measures of the deceived State.

4. LEGALITY OF THE ACT OF DECEPTION

Deception can also be approached from a different perspective, when the focus is not on the question of international responsibility and retaliatory measures of the deceived State but on the legal assessment of the very act of deception. In other words, it can be asked whether misleading a State into the commission of an internationally wrongful act against another State may itself amount to a breach of international law. Answering this question is a challenging undertaking. Deception is a matter not, per se, regulated by international law. Thus, it appears problematic to identify the rules of general international law that such conduct may be in contradiction with.[37]

International law contains a principle according to which a State must not knowingly allow its territory to be used for acts contrary to the rights of other States (*sic utere tuo* (no-harm) principle).[38] Max Huber even argued that a State bears the obligation to protect within its territory the rights of other States, including their integrity and inviolability.[39] This principle is recognized as a limitation of State sovereignty[40] and

35 ibid 287–288. Notably, the mere incitement is not unlawful in the law of State responsibility.

36 Derived responsibility may be relevant, for example, when a State knowingly provides its cyberinfrastructure to another State for the commission of a wrongful act by the latter may incur international responsibility for such assistance (*Tallinn Manual 2.0* (n 18) commentary to Rule 18, para 6). Derivative responsibility may also be relevant when cyber operation is conducted through computer networks infected and used remotely without the free will of the territorial State (e.g. botnets used against Estonia in 2007) – see François Delerue, *Cyber Operations and International Law* (CUP 2020) 307.

37 Deception may constitute a breach of obligations existing between the States concerned, such as multilateral or bilateral treaties on friendship and cooperation, but the existence and content of such obligations may vary. This contribution, therefore, focuses on obligations stemming from general international law.

38 *The Corfu Channel Case (United Kingdom of Great Britain and Northern Ireland v Albania)*, Judgment (1949) para 22.

39 Jutta Brunnée, 'Sic utere tuo ut alienum non laedas', in Rüdiger Wolfrum (ed) *Max Planck Encyclopedia of Public International Law* (Oxford University Press, updated March 2010) para 6, referring to Palmas Island Arbitration.

40 ibid paras 1, 4, 9.

in specific areas (in particular in international environmental law) as a distinct legal norm.[41] However, it is a question of whether this principle constitutes a stand-alone legal rule applicable in the cyber context, thus outside the context of international environmental law where the principle originally evolved.[42] Several authors claim that the obligation not to allow one State's territory to be used contrary to the rights of another State is now part of customary international law, even in relation to cyberspace.[43]

The *sic utere tuo* principle is often characterized as an obligation of due diligence nature.[44] Under the due diligence standard, a State is responsible if it knew about a cyber operation carried out from its territory, where the operation was contrary to the rights of another State and it failed to take feasible measures to prevent it.[45] That might in some situations ease the burden of the victim State to prove attribution as it would be enough to prove that the deceiving State's territory was used for the false-flag operation and that the deceiving State must have been aware of that and failed to prevent such event.[46]

However, the application of the no-harm principle raises several questions, such as the identification of the harm caused to the victim State and whether the harm originates from the territory of the deceiving State.[47] Can the causing of misattribution constitute a violation of the misled State's rights and amount to harm? Or is it the eventual damage caused to the third (victim) State by retaliatory measures adopted by the misled State that can be seen as the harm? If the former is correct, this harm can be understood as originating from the territory of the deceiving State since its false-flag attack caused it. However, if the latter is correct, it has to be asked whether the harm caused by the retaliatory measures is to be understood as originating from the territory of the deceiving State since it is also the result of the sovereign decision of the *deceived* State (although caused by deception).

Finally, one might also consider whether orchestration of the deception can or cannot amount to aggression. However, if the *false-flag attack that was used as a tool for deception* as such does not qualify as aggression due to the lack of use of armed force, then it can hardly be argued that the very act of *deceiving* a State into the commission of internationally wrongful conduct can be denoted as aggression either.[48]

41 ibid para 10; *Legality of the Threat or Use of Nuclear Weapons*, ICJ, Advisory Opinion (1996) para 29.
42 Brunnée (n 39) paras 9 and 16.
43 Russell Buchan, 'Cyberspace and the Obligation to Prevent Transboundary Harm' (2016) 21(3) Journal of Conflict and Security Law 429; Luke Chircop, 'A Due Diligence Standard of Attribution' (2018) 67(3) International & Comparative Law Quarterly 643; *Tallinn Manual 2.0* (n 18) Rules 6–7.
44 ibid (all).
45 Chircop (n 43) 650.
46 Delerue (n 36) 374.
47 The existence of harm is considered as one of the conditions of a violation of the no-harm principle.
48 Yoram Dinstein, 'Aggression', in Rüdiger Wolfrum (ed) *Max Planck Encyclopedia of Public International Law* (Oxford University Press, updated September 2015) para 16.

5. LEGAL EFFECTS OF MISATTRIBUTION CAUSED BY CYBER DECEPTION

One of the legal issues arising out of the misattribution caused by the deception is the legality of the retaliatory measures directed by the deceived State against the alleged wrongdoer. The legality of such measures depends on their legal character. If the deceived State responds with reciprocal measures, they may also qualify as a violation of the prohibition on intervention or the obligation to respect the sovereignty of other States.[49] However, retaliatory measures that are otherwise contrary to international obligations may be legal if they qualify as countermeasures or if they fulfil conditions of any of the (other) circumstances precluding wrongfulness.

It is the very first basic precondition of any countermeasure that it must be taken in response to a previous internationally wrongful act of another State and must be directed against that State.[50] This appears problematic in the case where technical evidence and intelligence information acquired by one State point towards another (victim) State as the wrongdoer, but in reality, the author of the false-flag cyber operation was different. Identification of the wrongdoer and attribution in the context of cyber operations is challenging because of the evidentiary and technical[51] peculiarities of cyberspace that make it possible to hide identity and leave false traces.[52] At the same time, the basic principle is that in bilateral disputes, the onus to establish responsibility lies on the injured State.[53] Together with the specific features of cyberspace, high demands are placed on the injured State in the process of attribution of harmful conduct.[54] Of course, retaliatory measures of the deceived State would also have to fulfil other requirements of countermeasures, namely a previous call upon the allegedly responsible State to fulfil its obligations and offer of negotiations.[55]

49 See the analysis above.

50 *Case Concerning the Gabčíkovo-Nagymaros Project (Hungary v Slovakia)*, Judgment (1997) para 83; *Tallinn Manual 2.0* (n 18) commentary to Rule 26, para 6.

51 For an overview of possible technical methods of attribution, see Massimiliano Albanese et al., 'Deceiving Attackers by Creating a Virtual Attack Surface' in Sushil Jajodia et al. (eds), *Cyber Deception: Building the Scientific Foundation* (Springer 2016) 150–151.

52 Robin Geiss and Henning Lahmann, 'Freedom and Security in Cyberspace: Shifting the Focus Away from Military Responses Towards Non-Forcible Countermeasures and Collective Threat-Prevention' in Katharina Ziolkowski (ed), *Peacetime Regime for State Activities in Cyberspace* (NATO CCD COE 2013) 625–626. At the same time, the basic principle is that in bilateral disputes, the onus to establish responsibility lies on the injured State. This places high demands on the injured State with respect to the process of attribution of a harmful conduct.

53 Draft Articles (n 22) commentary to Chapter V, para 8.

54 Carrying out in-depth investigation and gaining sufficiently convincing evidence may be time-consuming and take even years – see Geiss and Lahmann (n 52) 626. The more sophisticated the adversary, the longer the investigation may be to gain sufficient evidence – see Thomas Rid and Ben Buchanan, 'Attributing Cyber Attacks' (2015) 38 The Journal of Strategic Studies 1–2, 32. On the other hand, security and policy considerations may require prompt reaction to deter further cyber incidents.

55 For more details on function and preconditions of countermeasures in the domain of cyberspace, see Geiss and Lahmann (n 52) 628–644, and *Tallinn Manual 2.0* (n 18) Rules 20–25. Since countermeasures shall not affect the obligation to refrain from the threat or use of force, it would also be necessary to assess whether the destruction of a critical infrastructure in State C could not amount to the use of force. See *Case Concerning Military and Paramilitary Activities in and against Nicaragua (Nicaragua v United States of America)*, Judgment (1986) para 249; Draft Articles (n 22) art 50 para 1 letter a).

The question that is pertinent in the context of cyber deception is the relevance of the mistake of fact that led to misattribution. Mistake of fact plays a role in some areas of international law, such as international criminal law or the law of international treaties.[56] With regard to the law of State responsibility, the relevance of mistake of fact can be discussed in relation to the criteria establishing State responsibility, circumstances precluding wrongfulness and arguably in relation to the determination of reparation.[57]

When considering criteria for the establishment of international responsibility (breach of an international obligation and attribution), a mistake of fact does not play any role since these criteria are objective in nature[58] and any subjective considerations are not relevant.[59] Once a breach of an international obligation is established and is attributable, it is prima facie sufficient to establish responsibility.[60] Fault,[61] *culpa* or *dolus* of the organs in question is not required unless otherwise provided for by the primary norm in question.[62]

The same applies to the question of the relevance of mistakes of fact with respect to circumstances precluding wrongfulness. First, a mistake of fact does not constitute a stand-alone circumstance precluding wrongfulness recognized in international law, including in cyberspace. Errors or mistakes of fact are absent from the authoritative list of circumstances precluding wrongfulness in the International Law Commission's (ILC) Draft Articles on Responsibility of States.[63] Second, a mistake of fact also

56 Marko Milanovic, 'Mistakes of Fact When Using Lethal Force in International Law: Part I' (*EJIL: Talk!*, 14 January 2020) <https://www.ejiltalk.org/mistakes-of-fact-when-using-lethal-force-in-international-law-part-i/> accessed 27 January 2020, and the following parts.

57 Art 39 of the Draft Articles reflects the rule that conduct of the injured State that contributed to its damage should be considered. Naturally, this does not apply to the wilful conduct or negligence of the wrongdoer (deceived State) and this rule probably cannot be extended to the scenario where it is a wilful act of a third State that contributed to the injury.

58 There are some subjective elements in the realm of derivative responsibility discussed below (e.g. aid or assistance or coercion). See Crawford (n 22) 405.

59 ibid 61; Draft Articles (n 22) commentary to art 2, para 3; Mary Ellen O'Connell, *The Power & Purpose of International Law: Insights from the Theory & Practice of Enforcement* (OUP 2008) 248. For an explanation of a different approach, see Giuseppe Palmisano, 'Fault', in Rüdiger Wolfrum (ed) *Max Planck Encyclopedia of Public International Law* (Oxford University Press, updated September 2007) paras 6–14

60 Crawford (n 22) 61.

61 O'Connell (n 59) 248.

62 Crawford (n 22) 61; Draft Articles (n 22) commentary to art 2, para 3. For an explanation of a different approach, see Palmisano (n 59) paras 6–14.

63 Delerue (n 36) 228; O'Connell (n 59) 249; See also the argumentation in favour of an 'objectivist' approach in Marko Milanovic, 'Mistakes of Fact When Using Lethal Force in International Law: Part II' (EJIL: Talk!) <https://www.ejiltalk.org/mistakes-of-fact-when-using-lethal-force-in-international-law-part-ii/> accessed 27 January 2020. There are few opinions that a mistake of fact might qualify as a circumstance precluding wrongfulness, but often they are not further elaborated on. See for example *Second report on State responsibility, by Mr James Crawford, Special Rapporteur* (1994, A/CN.4/498 and Add. 1–4) para 262. An error has significant legal relevance in some other sub-fields of international law, for example as a ground for invalidity of international treaties – see Article 48 of the Vienna Convention on the Law of Treaties from 1969, or in the context of international criminal law – see for example art 32 para 1 of the Rome Statute of the International Criminal Court from 1998. See an overview of state practice on the relevance of mistake of facts in various sub-fields of international law in Milanovic (n 56) and the following parts.

cannot serve as a ground for the invocation of any of the established circumstances precluding wrongfulness since none of them requires an inquiry into a subjective element, such as fault[64] or error.[65] Specifically, with respect to countermeasures, the law of State responsibility is based on an objective standard.[66] A State resorting to countermeasures does so at its own risk and on the basis of its unilateral assessment of the situation. An incorrect assessment, including in the event of misattribution of malicious cyber operations,[67] may result in the commission of a wrongful act by the State resorting to countermeasures for which that State would be internationally responsible.[68] Interestingly, according to the ILC, there is no difference between countermeasures and other circumstances precluding wrongfulness in this respect.[69] Similarly, also in the context of self-defence, mistake or error of fact is irrelevant as the requirements for invoking self-defence are 'strict and objective'.[70]

Depending on the nature, timing and context, the illegality of actions of the deceived State taken against the alleged wrongdoer might in some situations be precluded by the reference to the state of necessity. The point is that actions taken with reference to the plea of necessity need not respond to an internationally wrongful act[71] and can violate the rights of non-responsible States.[72] However, necessity could only be invoked if it could be established that the reactive measure adopted by the misled State is the only means for the State to safeguard an essential interest (e.g. protection of an electric grid or critical infrastructure)[73] against a grave and imminent peril caused by the cyber threat. In light of the above-described objective standard, this can hardly be the case if the victim State is not at all involved in the false-flag attack. But one could imagine a different scenario where a computer network located in the victim State was exploited by the deceiving State and still remains an imminent threat to the critical infrastructure of the deceived State.[74]

Consequently, even when misattribution is caused by the wilful deception orchestrated by a deceiving State, conditions of international responsibility of the deceived State may be established due to their objective nature. Thus, it seems that rules of

64 O'Connell (n 59) 249.

65 In the context of self-defence, compare *Oil Platforms (Islamic Republic of Iran v United States of America)*, ICJ, Judgment (2003) para 73. Mistakes and involuntary acts may constitute unlawful use of force. See also Delerue (n 36) 305.

66 Draft Articles (n 22) commentary to art 49, para 3.

67 ibid art 4; *Tallinn Manual 2.0* (n 18) commentary to Rule 20, para 16.

68 Draft Articles (n 22) commentary to art 49, para 3; Delerue (n 36) 438.

69 Draft Articles (n 22) commentary to art 49, para 3; for more details on function and preconditions of countermeasures in cyberspace, see Geiss and Lahmann (n 52) 644–652.

70 *Oil Platforms (Islamic Republic of Iran v United States of America)*, ICJ, Judgment (2003) para 73. Mistakes and involuntary acts may constitute unlawful use of force. See Delerue (n 36) 305. See also considerations in Milanovic (n 56).

71 Chircop (n 43) 656.

72 *Tallinn Manual 2.0* (n 18) commentary to art 26, para 6.

73 ibid para 5.

74 Geiss and Lahmann (n 52) 646, referring also to national cyber security strategies. See also *Tallinn Manual 2.0* (n 18) commentary to Rule 26, paras 3–8 and Delerue (n 36) 348–350.

international responsibility (with some exceptions in Article 39 of the ILC's Draft Articles related to the determination of reparation once responsibility is established) do not contain any legal concepts that would take into account an excusable mistake caused by another State in the process of attribution. Various solutions to the consequences caused by wrongful attribution are of course possible if the deceived and the victim States reach a corresponding agreement (or the victim State does not invoke the responsibility of the deceived State). But that would require a constructive approach and clear establishment of the facts (including technical evidence), which might not be easy to achieve in some political contexts in combination with the complexity of the cybersphere.

6. CONCLUSION

Deception in cybersphere is a threat and a sophisticated challenge for international security. As this contribution has proven, it also constitutes a legal challenge. The analysis demonstrated that due to the lack of regulation of the issue in international law, it is problematic to identify international law rules that the deceptive action could violate. A possibly applicable principle, the no-harm rule, has been identified; however, its use appears problematic and deserves broader discussion. At the same time, the rules of international responsibility do not allow for the invocation of responsibility of a State that conducted deceptive action for the damage caused by the deceived State to the victim State. Moreover, a mistake of fact on the side of the misled State does not generally alleviate its responsibility for the retaliatory measures against the alleged wrongdoer, since it does not qualify as a distinct circumstance precluding wrongfulness, nor is it relevant for the invocation of any of the recognized circumstances precluding wrongfulness.

From the analysis, it seems apparent that a legal gap exists in international law. Rules of international responsibility address various forms of what is known as derived international responsibility, but deception does not fall into any of them. Legal consequences of misattribution in cyberspace are not sufficiently addressed in the doctrine and the topic appears to pose a challenge for the law of international responsibility. This outcome is not desirable since mistaken attribution may contribute to the destabilization of international peace and security.[75] Therefore, States carry a great burden to use all efforts to properly investigate and assess the background of cyber operations they face, and they must very carefully balance the need to

75 Delerue (n 36) 190.

effectively respond to them on the one hand and the legal and political consequences of misattribution on the other.[76]

The invocation of international responsibility of the deceived State is at the discretion of the victim State and may be settled by an agreement if the intent of the deceiving State is unveiled. However, in a situation when the two States' relations are not particularly friendly (something the deceiving State may be fully aware of when choosing the target of the false-flag attack), an agreement is less likely.

76 For an overview of State positions related to errors and mistakes in the non-cyber context (e.g. land-frontier and aerial incidents), see *'Force majeure' and 'Fortuitous event' as circumstances precluding wrongfulness: Survey of State practice, international judicial decisions and doctrine – study prepared by the Secretariat* (Extract from the Yearbook of the ILC 1978, Vol. II/1, A/CN.4/315) para 118 and following.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)

Military Data and Information Sharing – a European Union Perspective

Sebastian Cymutta*
Law Researcher
NATO CCDCOE
Tallinn, Estonia
sebastian.cymutta@ccdcoe.org

Marten Zwanenburg*
Professor of Military Law
University of Amsterdam and
Netherlands Defence Academy
Amsterdam, Netherlands
m.c.zwanenburg@uva.nl

Paul Oling*
PhD candidate Intelligence & Security
Netherlands Defence Academy
Breda, Netherlands
p.oling@mindef.nl

Abstract: The use of biometric data during and beyond military operations has become a top priority for the North Atlantic Treaty Organization (NATO) in recent years. But biometrics has also been relevant for European Union (EU)-led operations. The use of biometrics in multinational operations, particularly the sharing of biometric data, raises important legal questions. This is particularly the case for EU-led operations, which operate in the framework of an organization that has a strong focus on the protection of the right to privacy and on data protection.

This paper intends to address legal questions surrounding the use of biometric data for different purposes in the course of a multinational military operation, with a focus on EU-led operations. The article has a special emphasis on the sharing of biometric data, both between (EU member) States and between domains.

Keywords: *biometrics, Common Security and Defence Policy (CSDP), data protection, European Union law, fundamental rights, 'Ping + Ring'*

* The views expressed in this article are the authors' alone and do not reflect the official position of any organization they might be working for.

1. INTRODUCTION

The use of biometric data for the purpose of verification or identification has become widespread in today's society. Unlocking a smartphone via fingerprint or passing airport security using facial recognition are examples of biometric verification. Searching a biometric database to connect a latent fingerprint to a known criminal is an example of biometric identification.

Simultaneously, biometric technology is increasingly adopted in the area of military intelligence and security due to its potential to strip adversaries of the advantage of anonymity. Military adoption of biometrics has already led to its use in a number of multinational operations, such as during North Atlantic Treaty Organization (NATO)-led military operations in Afghanistan and operations in Iraq.[1]

The use of biometric systems in European Union (EU)-led missions is far less common and has received little attention so far. Considering the potential benefits of this technology in a military environment, it is likely that there will be a push for more widespread use within such missions. An example of this development is the EU-led operation EUNAVFOR MED IRINI. Established in 2020, its primary task is enforcing the United Nations arms embargo on Libya, to which end the mission may 'collect and store, in accordance with applicable law, personal data concerning persons involved in the carriage of such prohibited items related to characteristics likely to assist in their identification, including fingerprints'.[2]

This article focuses on the application of EU legislation concerning data protection to the sharing of biometric data within EU-led military missions. The use of biometrics is more prevalent in NATO-led operations than in EU-led missions. This is mainly a consequence of the fact that the United States, which is at the forefront of military use of biometrics, is part of NATO and not the EU. Unlike NATO, the EU is unique as an international organization, in that it has developed an extensive framework for the protection of data, including biometric data. This raises expectations concerning legal safeguards when using biometric systems in military missions led by the EU, especially with respect to the sharing of biometric data. Against this background, this article discusses how the EU data protection framework impacts the processing and sharing of biometric data in the context of EU-led missions. It may be noted, however, that the conclusions of this article may also be relevant for NATO operations, as many EU member States are also members of NATO and may be bound by EU data protection law when taking part in NATO operations.[3]

1 Annie Jacobsen, *First Platoon: A Story of Modern War in the Age of Identity Dominance* (Penguin 2021).
2 Council Decision (CFSP) 2020/472 of 31 March 2020 on a European Union military operation in the Mediterranean (EUNAVFOR MED IRINI) [2020] OJ L101, arts 2(6) and 4(5).
3 This is the case in any event for those EU and NATO member States that have domestic legislation that makes EU data protection law or parts thereof applicable to their armed forces. See section 3 below.

The article is structured as follows: after this introduction, section 2 provides a brief introduction to EU military missions and the use of biometric data. Section 3 gives an overview of the EU legal framework pertaining to data protection. This framework will be applied to biometric data sharing in the Common Security and Defence Policy (CSDP) domain in section 4. To provide some more insight into the concrete application of the relevant law, this section includes a discussion of several cases of such sharing. The article concludes with a number of final remarks.

Limitations on the use of biometric data and information sharing in EU-led missions may also follow from the application of the European Convention on Human Rights (ECHR), to which all EU member States are parties. Owing to space constraints, the application of the ECHR will not be addressed in this article.[4]

2. EU MILITARY MISSIONS AND THE USE OF BIOMETRIC DATA

A. The EU as a Military Actor

Since the new millennium, the EU has taken up a much more active role when it comes to military endeavours, starting in 2003 with the first EU military operation Concordia in what is now North Macedonia.[5] As of March 2022, the EU was conducting 7 military missions as well as 11 civil missions,[6] sometimes within (or near) the same region as military operations conducted by NATO.[7]

The need for a military endeavour usually arises in regions that are not only far less economically developed than the EU member States, but also are considerably more dangerous. Indeed, over the years, EU military missions have frequently been deployed to volatile regions. One example of this is the European Union Training Mission in Mali, which has seen a number of incidents.[8]

4 See for more on this topic e.g. Steven van de Put and Marten Zwanenburg, 'Military Use of Biometrics and the Right to Private Life in Article 8 ECHR' (working title) [2022] NL ARMS (forthcoming).

5 Council Joint Action 2003/92/CFSP of 27 January 2003 on the European Union military operation in the Former Yugoslav Republic of Macedonia [2003] OJ L 34; see generally on the development of EU military endeavours Sabine Mengelberg, 'Permanent Change: the Paths of Change of the European Security Organizations' (PhD thesis, Leiden 2021).

6 See for more information, 'Military and civilian missions and operations' (European Union, 5 March 2019) <https://www.eeas.europa.eu/sites/default/files/eu_csdp-missions-and-operations_2021-10.pdf> accessed 12 April 2022.

7 For example: EULEX in Kosovo and KFOR / EUAM Iraq and NATO Mission Iraq.

8 See e.g. Deutsche Welle, 'Gunmen Attack Bamako Base of EU Military Training Mission in Mali' (Berlin, 21 March 2016) <https://www.dw.com/en/gunmen-attack-bamako-base-of-eu-military-training-mission-in-mali/a-19132542> accessed 2 March 2022; Associated Press, 'EU Training Mission Comes under Attack in Mali' (New York City, 14 February 2019) <https://apnews.com/c375c2a0628b43d286c71ef06f5fb89f> accessed 2 March 2022.

B. The EU's Common Foreign and Security Policy

EU military missions are conducted within the intergovernmental[9] framework of the Common Security and Defence Policy (CSDP), which is a subcategory of the EU's Common Foreign and Security Policy (CFSP). The legal basis for the CFSP is found in Title V of the Treaty on European Union (TEU). Articles 42(1) and 42(3) TEU provide the EU with an operational capacity drawing on civilian and military assets of the member States. The EU may use these assets for 'missions outside the Union for peace-keeping, conflict prevention and strengthening international security in accordance with the principles of the United Nations Charter'.[10]

Whereas Article 42 TEU refers only to 'missions', it is common to refer to 'operations' when talking about executive or military endeavours and to 'missions' when talking about non-executive endeavours. For ease of reference, this article will use the term 'mission' to refer to both operations and missions.

An important assumption in this article is that biometric data collected by a participating State in an EU-led mission is 'owned' by that State, and not by the EU. This is how the issue of 'ownership' is approached in NATO missions and we do not see a need to take a different approach in the context of EU-led missions.[11]

C. Biometrics and the 'Ping & Ring' Concept

'Biometrics' or 'biometric recognition' is defined as the automated recognition of individuals based on their biological and behavioural characteristics.[12] It uses the physical, physiological and/or behavioural characteristics of individuals to recognize them.[13] Examples of such characteristics are face topography, hand topography, finger topography, iris structure, vein structure of the hand, voice, gait, and DNA.[14] These characteristics are unique, which makes them very suitable for recognizing persons.[15]

A 'biometric system' is defined as a system for the purpose of biometric recognition of individuals based on their behavioural and biological characteristics.[16] It is essentially a pattern recognition system that operates by acquiring biometric data from an

9 Grabitz/Hilf/Nettesheim/Kaufmann-Bühler EUV art 42 paras 13, 14.
10 Treaty on European Union (Maastricht Treaty) art 42(1).
11 Hence this paper does not discuss the data protection rules pertaining to EU bodies and institutions and therefore does not consider Regulation (EU) 2018/1725, seeing that this body of law explicitly excludes itself from being applicable to entities created according to articles 42(1), 43, 44 TEU.
12 ISO/IEC International Standard 2382-37, *Information Technology – Vocabulary – part 37: biometrics* 2 (2012).
13 See for an extensive description of biometrics inter alia Nancy Y Liu, *Bio-Privacy: Privacy Regulations and the Challenge of Biometrics* (1st edn Routledge, 2012) 29–59.
14 For additional characteristics see William H Boothby, 'Biometrics' in William H Boothby (ed) *New Technologies and the Law in War and Peace* (Cambridge University Press 2019), 192.
15 'Recognizing' is used here as a term encompassing verification and identification as defined below.
16 ISO/IEC International Standard 2382-37 (n 12).

individual, extracting a feature set from the acquired data, and comparing this feature set against the template set in the database.[17]

The use of biometrics in multinational military operations may take the shape of biometric 'ping & ring networks'. The 'ping & ring' concept involves biometric data queries which, when matched in certain databases, yield a reference number and a point of contact for follow-on bilateral action.[18] Personnel from one State may send biometric data to another State and request that the other State use the data to run a query in their biometric database. The query does not give them direct access to the actual biometric data in the database, or to the biometric data subject's biographic or contextual data. If there is a match, the State that made the request may ask for more information. In other words, a query in a ping & ring network involves the requesting State sharing biometric data. The sharing of data from the database by the requested State may follow on a bilateral basis if there is a match.

D. The (Potential) Role of Biometric Data Sharing

Military organizations increasingly face adversaries that 'simultaneously and adaptively employ a fused mix of conventional and improvised weapons, irregular tactics, terrorism and criminal behaviour in the battlespace to obtain their political objectives'.[19] A striking example is the insurgents' use of Improvised Explosive Devices (IEDs), the signature weapon of recent asymmetric conflicts in Iraq, Afghanistan, Mali and Syria. Unexpected external shocks, such as the IED threat faced in Iraq and Afghanistan, are 'fertile ground for innovation'.[20] To counter the threat of IEDs, military organizations utilize emerging technologies pertaining to the intelligence and security domain.

Initially, most counter-IED (C-IED) efforts focused on defensive security technologies, such as improved armoured plating and tactics, techniques, and procedures to detect IEDs.[21] Over time, C-IED efforts became more proactive with an emphasis on intelligence.[22]

The use of biometrics by military organizations is an example of both strategies. Biometric technology was deployed in Kosovo as part of the counter-intelligence activities to grant authorization to individuals accessing military bases.[23] In Iraq and

17 Anil K Jain, Arun Ross and Salil Prabhakar, 'An Introduction into Biometric Recognition' (2004) 14 IEEE Transactions on Circuits and Systems for Video Technology 4, 5.

18 Victor Morris, 'Identity and Biometrics Enabled Intelligence (BEI) Sharing for Transnational Threat Actors' (2016) Small Wars Journal.

19 Frank Hoffman, '"Hybrid Threats": Neither Omnipotent Nor Unbeatable' (2010) 54 Orbis 441.

20 Adam Grissom, 'The Future of Military Innovation Studies' (2006) 29 Journal of Strategic Studies 905.

21 David W Barno and Nora Bensahel, *Adaptation under Fire: How Militaries Change in Wartime*. Bridging the Gap (Oxford University Press 2020).

22 Theo Farrell, Frans PB Osinga and James A Russell, *Military Adaptation in Afghanistan* (Stanford University Press, 2013).

23 Jacobsen (n 1) 48.

Afghanistan, coalition forces increasingly relied on biometrics as an offensive tool, supporting biometrics-enabled intelligence. Biometrics were used to link persons to times, locations, groups and activities, while simultaneously providing a means to detect and identify them in the future. Forensically exploiting IED components or remnants, documents and electronic data carriers allowed for the identification and disruption of networks of individuals supporting the IED threat.

As these networks are not restricted by artificially defined operational areas within a theatre, exchanging biometric data within a multinational mission is vital. Or as described by Arquilla et al., 'it takes a network to defeat a network'.[24] This intense cooperation during a multinational mission may lead to isomorphism between military organizations, where organizations try to emulate one another.[25]

However, using a solely deterministic approach to the introduction of biometrics does not adequately take into account the influence of societal acceptance of a technology on its use by military organizations. The acceptance of the use of biometrics at home shapes the approval for military use abroad. Moreover, the way a nation allows its military to use biometrics shapes the way technical interoperability between military organizations is achieved. For example, legal caveats for data retention – when applicable – must be made part of the technical interoperability, accompanying exchanged biometric data from the cradle to the grave. As the United States may be considered a pacesetter for the military application of biometrics, their societal acceptance of the technology and its legal safeguards resulting from that is reflected in current military-technical standards on interoperability.

3. THE LEGAL FRAMEWORK OF THE EU PERTAINING TO DATA PROTECTION

A. Primary Law Implications

Article 39 TEU is a statutory source for CFSP-specific rules on data protection.[26] It refers to Article 16 of the Treaty on the Functioning of the European Union (TFEU) and tasks the Council with the adoption of legislation laying down the rules relating to the protection of individuals with regard to the processing of personal data by the member States when carrying out activities which fall within the scope of the chapter in the TEU dealing with the CFSP, and the rules relating to the free movement of such data.

24 John Arquilla and David Ronfeldt, *Networks and Netwars: The Future of Terror, Crime, and Militancy* (RAND Corporation 2001).

25 Theo Farrell and Terry Terriff (eds), *The Sources of Military Change: Culture, Politics, Technology, Making Sense of Global Security* (Lynne Rienner Publishers 2002).

26 Calliess/Ruffert/Kingreen EU-Vertrag (Lissabon) (2022) art 39 para 1; Pechstein/Nowak/Häde, Frankfurter Kommentar EUV/GRC/AEUV/von Heinegg EUV (2017) art 39 para 2.

Article 39 highlights for the first time[27] that the CFSP is not a data protection-free area of EU policy and action, drawing on the material provisions of Article 16 TFEU. While Article 16 is the most prominent provision regarding data protection in EU primary law, Article 8 of the EU Charter of Fundamental Rights (CFR) further emphasizes the concept of data protection as a human right.[28] While technically not considered part of the primary law of the EU, the CFR is recognized by the EU and has the same legal force as the Treaties.[29] Article 8(1) CFR provides that 'everyone has the right to the protection of personal data concerning him or her'. In its second subsection, it is further stipulated that 'such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law and that everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified'. The objective of this fundamental right, it is suggested, is to protect an individual's control over personal data.[30]

Article 8 CFR is based in part on Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (Directive 95/46/EC), the forerunner of Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation (GDPR)).[31] It is submitted that for this reason, Directive 95/46/EC and the GDPR as its successor may be instructive in establishing what Article 8 CFR requires.[32]

Since Article 51 CFR limits the Charter's scope of application to the member States only when implementing Union law, it begs the question whether the armed forces of a member State participating in a CSDP mission would qualify as 'implementing Union law'. Naert has rightly argued that this is indeed the case. This is because 'implementing Union law' encompasses situations in which member States implement

27 Thomas Ramopoulos, 'Article 39', in Manuel Kellerbauer, Marcus Klamert and Jonathan Tomkin (eds), *The EU Treaties and the Charter of Fundamental Rights* (Oxford University Press 2019) 1159.

28 Charter of Fundamental Rights of the European Union [2012] OJ C326/02.

29 See art 6(1) TEU.

30 Tobias Lock, 'Article 8 CFR', in Kellerbauer et al (n 27) (2019) 2123.

31 Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data [1995] OJ L281; Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119.

32 See Herke Kranenborg, 'Article 8', in Steve Peers, Tamara Hervey, Jeff Kenner and Angela Ward (eds), *The EU Charter of Fundamental Rights: A Commentary* (2nd edn, Hart 2014) 223, 247.

Council acts setting up EU military operations, as such decisions are legal (albeit not legislative) acts under EU law.[33]

Article 52 CFR allows for derogations from the rights set out in the CFR. However, any limitation must be provided for by law, and respect the essence of those rights and freedoms. Subject to the principle of proportionality, limitations may be made only if they are necessary and genuinely meet objectives of general interest recognized by the Union or the need to protect the rights and freedoms of others.

B. Secondary Law Implications

1) The Need for Concretization

While EU primary law contains some strong fundamental rights with regard to data protection, these rights are largely dependent on being substantiated by EU secondary law. This holds true especially for the CFSP domain. Even though the fundamental right of data protection is applicable within the CFSP, the primary law of the EU recognizes that the peculiarities of the CFSP (and especially the CSDP) warrant a different kind of data protection regime than the one established on the basis of Article 16(2.1) TFEU for the civilian sector.[34] It is Article 16(2.2) TFEU that recognizes the need for specific rules laid down by Article 39 TEU for the CFSP. Hence, Article 39 TEU in conjunction with Article 16 TFEU obliges[35] the Council to enact a special data protection regime that takes into account the unique circumstances of endeavours undertaken within the CFSP.[36] Unlike in Article 16 TFEU, the European legislator does not enjoy freedom of choice with regard to the secondary law instruments provided in Article 288 TFEU. Article 39 TEU calls for a decision in the sense of Article 288(4) TFEU, which 'shall be binding in its entirety'.

2) Filling the Gap

Despite the above-mentioned obligation to do so, the Council has adopted no decision so far. Since a gap in secondary legislation is undesirable, mitigation might be achieved by applying existing EU data protection legislation.[37]

33 Frederik Naert, 'Shared Responsibility in the Framework of the European Union's Common Security Defense Policy Operations' in André Nollkaemper and Ilias Plakokefalos (eds), *The Practice of Shared Responsibility in International Law* (Cambridge University Press 2016); Carmen Márquez Carrasco, 'Human Rights in the EU's Common Security and Defence Policy', in Jan Wouters, Manfred Nowak, Anna-Luise Chané and Nicholas Hachez (eds), *The European Union and Human Rights: Law and Policy* (Oxford University Press 2020) 408, 415.

34 Most notably the GDPR (n 31) and Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC [2018] OJ L295.

35 Streinz/Regelsberger/Kugelmann EUV (2018) art 39 para 1.

36 Pechstein/Nowak/Häde, Frankfurter Kommentar EUV/GRC/AEUV/von Heinegg EUV (2017) art 39 para 3.

37 Grabitz/Hilf/Nettesheim/Kaufmann-Bühler EUV (2021) art 39 para 5.

The GDPR is the most influential secondary law act in the area of data protection. However, Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data is lex specialis in relation to the GDPR, as far as the protection of natural persons specifically with regard to the processing of personal data for law enforcement purposes is concerned.[38]

What the Regulation and the Directive have in common is that their scope of application does not extend to data processing 'in the course of an activity, which falls outside the scope of Union law'.[39] In addition, the GDPR does not apply to data protection activities by the member States 'when carrying out activities which fall within the scope of Chapter 2 of Title V of the TEU'.[40] Consequently, neither the GDPR nor Directive 2016/680 can provide a fully applicable data protection regime for CSDP-missions. Yet in this context, it is important to note that a number of EU member States have unilaterally extended the application of the GDPR[41] or enacted comparable legislation[42] to govern the activities of their armed forces. Hence, it is fair to say that at least some of the provisions of the GDPR were considered to be a good fit for the needs of these States, when trying to establish a data protection regime for their militaries. Therefore, this paper will approach the idea of applying the GDPR / Directive 2016/680 by way of analogy, when assessing biometric data sharing within the CSDP-domain.

4. BIOMETRIC DATA SHARING IN THE CSDP DOMAIN

In this section, the framework set out in section 3 will be applied to biometric data sharing in the CSDP domain. Subsections C to E discuss several cases of such sharing, providing insight into the concrete application of the relevant law.

38 Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA [2016] OJ L119.

39 See Directive 2016/680 (n 38), art 2(3)(a) and GDPR (n 31) art 2(2)(a).

40 See GDPR (n 31), art 2(2)(b).

41 Sebastian Cymutta, 'Biometric data processing by the German armed forces during deployment' (2021) CCDCOE, 7–8.

42 For example, the Netherlands has implemented the Uitvoeringswet Algemene Verordening Gegevensbescherming and Regeling Gegevensbescherming Militaire Operaties.

A. Utilizing the GDPR

As it is directly applicable to the CSDP, Article 8 CFR will be the primary reference point when assessing the legality of biometric data processing,[43] which includes sharing, in EU military missions.

It follows from the second paragraph of Article 8 CFR that the sharing of biometric data requires either the consent of the person concerned or some other legitimate basis laid down by law. Consent must be understood as informed consent.[44] If consent has not been given, sharing is still possible if there is another legitimate basis, but only if this is laid down by law. As was already stated, Article 8 CFR is based in part on Directive 95/46/EC, the forerunner of the GDPR. It is submitted that for this reason, Directive 95/46/EC and the GDPR as its successor may be instructive in establishing what Article 8 CFR requires.[45]

For the purposes of this paper, Article 9 GDPR is of particular importance, as it contains specific rules on the processing of special categories of personal data, including biometric data, for the purpose of uniquely identifying a natural person.[46] According to this provision, the processing of biometric data is in general forbidden, except in the cases provided for in Article 9(2) GDPR. Of particular interest for CSDP missions is Article 9(2)(g), according to which processing could be allowed 'if it is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject'. This very high standard is reflective of the sensitive nature of biometric data, and while Article 9 does not directly apply to CSDP missions, it is reasonable to assume that a similarly high standard applies under Article 8 CFR.

B. Legal Principles Influencing Biometric Data Sharing

Legislation allowing for the processing of data must lay down clear and precise rules governing the scope and application of measures, and its application should be foreseeable to persons subject to it.[47] It should also impose minimum safeguards concerning duration, storage, access for third parties, procedures for preserving the integrity and confidentiality of data and for its destruction, as well as sufficient

43 Processing is very broadly understood as to include 'any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction', see GDPR (n 31), art 4(2).

44 Norbert Bernsdorff, 'Artikel 8' in Jürgen Meyer (ed), *Charta der Grundrechte der Europäischen Union* (3rd edn, Nomos 2010) para 21.

45 See Kranenborg (n 32) 247.

46 GDPR (n 31) art 9(1).

47 GDPR (n 31) recital 41.

guarantees against the risk of abuse and arbitrariness.[48] In particular, the automatic processing of data increases the need for such safeguards.[49]

It follows from Article 8 CFR that the sharing of data may only be done for specified purposes. Although this is not explicitly stated, it is submitted that this includes the applicability of the 'purpose limitation principle' or 'finality principle', which is one of the cornerstones of data protection.[50] This principle, which is also set out in Article 5(1)(b) of the GDPR, requires that data should not be further processed in a way that is incompatible with the purposes for which it was originally processed (subject to limitations allowed under Article 52 CFR).

The principles of necessity and proportionality are also fundamental elements of data protection. They are referred to in Article 5(1)(c) of the GDPR. This provision states that personal data should be 'adequate, relevant and limited to what is necessary for the purposes for which they are processed ("data minimisation")'. Necessity requires that if less intrusive means can achieve the same purpose, personal data may not be processed.

Proportionality requires an assessment of the impact of the right to personal data protection against the constitutional value it aims to achieve. To ascertain whether interference is disproportionate, it is necessary to consider how the right to data protection will be restricted – for instance, the type of data that will be processed and whether it involves specially protected data – and the safeguards in place.[51]

Transparency is another important principle of data protection. It is reflected in Article 5(1)(a) GDPR, which provides, inter alia, that personal data shall be processed 'in a transparent manner in relation to the data subject'. The transparency of data processing is part of what constitutes 'fair' processing, as referred to in Article 8 CFR.[52]

C. Base Access and Data Transfer

As far as the authors are aware, EU missions have not yet used biometrics in the context of base access. However, this could change in the future, as this has become a common methodology in the last decade. It has also been suggested that in order to be truly effective, a biometric access control system should be able to exchange data with the biometrics systems being used by operational forces, and that an access control

48 ibid; see also Lock (n 30), referring to *S and Marper v UK* App nos 30562/04 and 30566/04 (ECtHR, 4 December 2008) para 99.

49 Joined Cases C-293/12 and C-594/12, *Digital Rights Ireland Ltd v Minister for Communications, Marine and Natural Resources and Others and Kärntner Landesregierung and Others* [2014] para 55.

50 See Kranenborg (n 32) 247.

51 Antonio Troncoso Reigada, 'The Principle of Proportionality and the Fundamental Right to Personal Data Protection: The Biometric Data Processing' (2012) 17 Lex Electronica 2, 18.

52 Kranenborg (n 32) 254.

system should be able to use the same biometric watch list being used by forces in the field.[53]

It follows from the above that such sharing of data would require explicit consent or another legitimate basis laid down in law. To the extent that consent for sharing their data would be requested from persons being enrolled in the biometric system, such consent must be freely given. It may be questioned, however, whether this is possible under the circumstances. Recital 43 of the GDPR provides that 'in order to ensure that consent is freely given, consent should not provide a valid legal ground for the processing of personal data in a specific case where there is a clear imbalance between the data subject and the controller, in particular where the controller is a public authority, and it is therefore unlikely that consent was freely given in all the circumstances of that specific situation'.

Such a power imbalance is clearly present in relations between a military operation and local personnel seeking access. As to the presence of another legitimate basis, such a basis must be laid down in law. This is arguably the case for member States that have domestic legislation concerning biometrics in military operations, such as the Netherlands and Germany.[54] For States that do not have such national legislation, the question is whether an international legal basis for the operation is sufficient. This could be either the EU Council Decision that forms the basis for the mission under EU law, or the legal basis under general international law, for example, a document setting out the consent of the host State to the mission or a Resolution of the United Nations (UN) Security Council under Chapter VII of the UN Charter. Arguably, the wording of Article 8 CFR does not exclude the possibility that such sources (EU or international) of law satisfy the 'laid down in law' criterion. However, such instruments would need to meet high standards that set out clear and precise rules governing the scope and application of measures and impose minimum safeguards concerning duration, storage, access for third parties, procedures for preserving the integrity and confidentiality of data and for its destruction as well as sufficient guarantees against the risk of abuse and arbitrariness.[55] It seems clear that the mere granting by the UN Security Council of the power to 'use all necessary means' to achieve the mission's mandate, for example, would not meet these requirements.

The 'transparency principle' sets high standards for the communication between the member State collecting biometric data for controlling base access and the data subject. At the very least, it requires that information be made available in a language that the local population understands. Furthermore, it may be wondered whether it is feasible in a CSDP mission to provide data subjects with access to their personal data, which is required by the transparency principle.

53 William C Buhrow, *Biometrics in Support of Military Operations: Lessons from the Battlefield* (1st edn, CRC Press 2017) 49–50.
54 See section 3B above.
55 See section 4B.

D. Biometric Data Sharing in Theatre

The 'purpose limitation principle' requires that the transfer of biometric data be limited to the purposes for which the data was originally collected. Arguably, this would allow the sharing of data collected in the context of base access with other States within the mission, as long as the transfer was for the purpose of ensuring the protection of the personnel of the mission, if this was defined as the purpose for collecting biometric data in the context of controlling base access. Protection of the personnel of the mission is a vital requirement for the mission to be able to fulfil its mandate.

With regard to the sharing of biometric data with States outside the mission, the case-law of the European Court of Justice (ECJ) with regard to the EU–Canada Passenger Name Record agreement suggests that it will be very difficult to find an adequate legal basis for sharing outside the mission. In that case, the ECJ rejected the grounds of 'protection of public security against terrorism and serious transnational crime' as a legal basis for sharing sensitive personal data.[56]

It has been argued that the general protection of personal data laid down in Article 8(1) CFR places limits on the transfer of data to third countries.[57] It is noted that this article assumes that biometric data is 'owned' by States when personnel of the State concerned participate in a mission led by an international organization. Consequently, the limits referred to above also apply to the transfer of data between (States participating in) EU-led missions and (States participating in) another mission, such as those led by NATO. To understand what the sharing of data with another State requires, it is informative to look at how this has been operationalized in the GDPR. Under that regulation, transfer of personal data to another State may only occur where that State affords 'adequate protection' of such data. This means that the protection guaranteed in the third country concerned must be 'essentially equivalent' to that under EU law.[58] The legal order of the third country must effectively protect the right to personal data, which includes clear limits to interferences with such data – concerning access, use, and so on – and procedural safeguards in place.[59]

According to the ECJ, the transfer of 'sensitive data' – such as racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership – 'requires a precise and particularly solid justification, based on grounds other than the protection of public security against terrorism and serious transnational crime'.[60] In the case in which the ECJ set out this standard, concerning an envisaged agreement

56 Opinion 1/15, *Transfer of Passenger Name Record data from the European Union to Canada* [2017] ECLI:EU:C:2017:592, para 165.
57 Lock (n 30) 2126.
58 Case C-362/14, *Maximillian Schrems v Data Protection Commissioner* [2015] ECLI:EU:C:2015:650, para 73.
59 ibid paras 79–98.
60 Opinion 1/15, *Transfer of Passenger Name Record data from the European Union to Canada* (n 56) para 165.

between the EU and Canada, biometric data was not involved. If it had been, in view of the special nature of biometric data, it is highly likely that the ECJ would have applied similarly high standards to the sharing of such data with another State.

Finally, Article 8(3) CFR requires the existence of a surveillance authority in each member State. The purpose of such an authority is to ensure the effectiveness and reliability of monitoring compliance with the law on the processing of personal data. Its aim is to strengthen the protection of individuals.[61]

In conclusion, based on an interpretation of Article 8 CFR using the GDPR and Directive 2016/680 by analogy, the sharing of biometric data in theatre requires establishing many safeguards. It is submitted that the ping & ring concept discussed before may allow for the integration of many of those safeguards.

E. Military-Civilian Data and Information Sharing

Another challenge would be the sharing of biometric data with other EU authorities. Exceptions aside, the purposes for which those authorities would use the data would most likely not be related to the purposes of the EU mission in the context of which the data was collected.

An example is EUNAVFOR MED IRINI,[62] which is currently operating in the South Central Mediterranean Sea under a mission mandate that aims to contribute to the disruption of the business model of human smuggling and trafficking networks and, not least of all, enforcing a UN arms embargo imposed on Libya.[63] To facilitate this, IRINI is allowed to collect personal data (including biometric data) and store it.[64]

Without negating the military character of IRINI, it is understood that the mandate is closely aligned to law enforcement efforts and that the biometric data collected during the mission is of interest and importance for EU law enforcement agencies. Hence, the EU Council Decision establishing IRINI provides for the transfer of this data to the relevant law enforcement authorities of member States and to competent Union bodies in accordance with applicable law.[65]

In the absence of a Council Decision detailing the (technical) rules relating to the processing (sharing) of personal data within the CFSP, the present reference to the 'applicable law' allows for a fallback to Article 8 CFR and the application by analogy of the provisions of the GDPR and Directive 2016/680. Arguably, for missions like IRINI, which have many characteristics of law enforcement operations, Directive

61 Case C-362/14, *Maximillian Schrems v Data Protection Commissioner* (n 58) para 41.
62 Council Decision (CFSP) 2020/472 (n 2); IRINI is the successor-mission to EUNAVFOR MED Sophia, which was established by Council Decision (CFSP) 2015/778 of 18 May 2015 on a European Union military operation in the Southern Central Mediterranean (EUNAVFOR MED) [2015] OJ L122/31.
63 ibid art 1.
64 ibid arts 2(6) and 4(5).
65 ibid.

2016/680 would be the most appropriate analogy. However, as was mentioned above, several EU member States have made the GDPR applicable to military operations through domestic legislation. This suggests that they consider the GDPR as providing an appropriate basis for the regulation of the use of biometrics by military missions. This is why the remainder of this section will refer to the GDPR.

Under the GDPR regime, while the collection of fingerprints would have to be measured against the underlying intention of Article 9 GDPR, the collection and storage could be considered legal under Article 9(2)(g).[66] It can be argued that the investigation and prosecution of persons involved in arms smuggling to Libya and human smuggling is a 'substantial public interest'. In that case, processing of biometric data would be allowed, provided that suitable and specific measures to safeguard the fundamental rights and the interests of the data subject have been taken. As there is no public information regarding how the processing of data collected in the context of IRINI takes place in general, it cannot be determined whether this requirement has been met.

However, that leaves the question of whether or not this biometric data can legally be shared with 'competent EU bodies' in the light of the principle of purpose limitation, as discussed above. It is submitted that a transfer of personal data from IRINI to another EU institution would constitute 'further processing' in the sense of Article 6(4) GDPR, and hence would have to be measured against a high standard. In practice, IRINI has concluded a Working Arrangement with the European Border and Coast Guard Agency (Frontex), detailing their cooperation[67] and focusing on 'cross-border crime such as arms trafficking and the disruption of the human smuggling model and trafficking networks'.[68]

This Working Arrangement details the Exchange of Information (and personal data) between IRINI and Frontex in section 5, where it provides for the (analogous) applicability of the GDPR and Article 8 CFR. This underlines the argument that in the absence of a Council Decision detailing data protection rules especially for the CFSP, referencing Article 8 CFR and the GDPR is a suitable workaround.

Therefore, this article argues that while the transfer of biometric data from IRINI to

66 Cymutta (n 41) 9.

67 A similar Working Agreement has been concluded between EUNAVFOR MED Sophia and EUROPOL, see 'Working Arrangement establishing cooperative relations between EUNAVFOR MED Operation Sophia and Europol' (EUROPOL, 16 January 2016) <https://www.europol.europa.eu/partners-collaboration/agreements/working-arrangement-establishing-cooperative-relations-between-eunavfor-med-operation-sophia-and-europol> accessed 2 March 2022.

68 See 'Working Arrangement between The European Border and Coast Guard Agency (Frontex) and EUNAVFOR MED IRINI' (Frontex, 18 January 2021), s 2(2), the conclusion of this kind of Working Agreements is foreseen in art 68(1)(j) of Regulation (EU) 2019/1896 of the European Parliament and of the Council of 13 November 2019 on the European Border and Coast Guard and repealing Regulations (EU) No 1052/2013 and (EU) 2016/1624 [2019] OJ L295; the predecessor mission 'Sophia' contained a corresponding provision in Council Decision 2015/778 (n 69) art 8(3).

Frontex is permissible on the basis of the instruments in place, its legality has to be measured against standards similar to the strict standards of Article 9(2)(g) GDPR.

5. CONCLUSION

This article discussed how the EU data protection framework impacts the processing and sharing of biometric data in the context of EU-led missions. In the absence of a decision by the Council as required by Article 39 TEU, the data protection regime that applies to such missions is not clearly defined. Understanding and developing the data protection regime pertaining to EU military missions is important for those carrying out such missions, and also to foster legal interoperability within a CSDP mission and beyond. As an example of interoperability beyond CSDP missions, there is likely to be increased cooperation between the EU and NATO in the future.

Whether it be the exchange of personal data between the member States, the enrolment of locally employed persons for granting base access or the sharing of personal data with bodies outside of the CSDP mission, all of these actions concern the (universal) right of data protection as it is provided for by the primary law of the EU.

While a comprehensive Decision of the Council detailing the rules for the processing of personal data in the CFSP is still missing, this paper showed that an acceptable standard of data protection (procedure) could be based on Article 8 CFR and the analogous application of certain provisions of the GDPR. Further legal substance would be given to these standards through agreements with the entities with which biometric data is shared, either through agreements with States or through 'Working Arrangements' with EU institutions. While this legal workaround appears to live up to practical needs, negotiating one or more (Working) Arrangements for each CSDP mission, potentially with several different States and EU institutions,[69] appears likely to be time-consuming and would potentially lead to new obstacles regarding legal interoperability. Seeing that EUNAVFOR MED IRINI recently struck another Working Arrangement with the EU Border Assistance Mission in Libya,[70] there is also a danger of the legal landscape fragmenting further. A comprehensive Decision on the basis of Article 39 TEU could not only eliminate the need for these kinds of Working Arrangements but could also provide adequate guidance for sharing personal data collected by CSDP missions with other security regimes like NATO or the UN, effectively contributing to international legal interoperability.

69 In addition to Frontex, it is also conceivable that European Union Agency for the Operational Management of Large-Scale IT Systems in the Area of Freedom, Security and Justice (eu-LISA) could be interested in receiving personal data collected during EU military operations.

70 See EUNAVFOR MED operation IRINI, 'Best practices, information & integrated approach. Operation EUNAVFOR MED IRINI signs a working arrangement with EUBAM Libya' (European Union, 7 August 2021) <https://www.operationirini.eu/best-practices-information-integrated-approach-operation-eunavfor-med-irini-signs-working-arrangement-eubam-libya/> accessed 2 March 2022 (as of March 2022 the text of the working arrangement was not available.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

Cyber Threats Against and in the Space Domain: Legal Remedies

Seth W. Dilworth*
Space Law Attorney
Operations and International Law
United States Air Force, Pentagon
Washington, D.C., United States
Seth.Dilworth.1@us.af.mil

D. Daniel Osborne*
Space & Operational Law Attorney
National Security Law Division
United States Army, Pentagon
Washington, D.C., United States
david.d.osborne.mil@army.mil

Abstract: Connecting traditional military domains (land, sea, and air), cyber and space domains are critical in the modern defense of worldwide assets. These domains leverage evolving technologies and international partnerships to further the national security interests of cooperating nations. International and domestic laws governing cyber and space assets rely on separate and distinct legal frameworks, effectively creating legal silos within each domain. Regardless, application of some key provisions of the Outer Space Treaty overlap with cyber operations and should be applied in the cyber threat context – state responsibility and liability. These provisions provide the framework to determine state liability under specific circumstances. Unique to the space domain, the Outer Space Treaty requires state responsibility for "national activities in outer space," a term undefined in the treaty. In accordance with the treaty provisions, states provide that responsibility through licensing regulations, statutes, oversight of launches, on-orbit activity, and other space-related conduct. Governing state liability, the liability provision and the subsequent Liability Convention render a state liable even after a state has transferred ownership of a satellite. Cyber operations conducted against space objects further complicate the legal remedy process, but should not preclude application of traditional space law. While the writers of the Outer Space Treaty could not have considered cyber operations as part of these provisions, these responsibility and liability provisions today partially bridge the gap between the legal frameworks mentioned previously. This paper argues legal practitioners can and should apply the responsibility and liability provisions to cyber operations and threats against space objects. Additionally, states can further bridge the gap between domains by addressing cyber operations in domestic space law. These applications modernize

* Unless otherwise noted, the conclusions expressed herein are solely those of the authors writing in their personal capacity. They are not intended and should not be thought to represent official ideas, attitudes, or policies of any agency of the United States Government, including the United States Army, United States Air Force, United States Space Force, or Department of Defense. The authors have used only information available to the public in the researching and presentation of this work.

the law and reflect the true reality of multi-domain assets and the technological reliance on space and cyber domains.

Keywords: *responsibility, liability, outer space, cyber, satellites, Outer Space Treaty*

1. INTRODUCTION

The space and cyber domains connect traditional military domains (land, sea, air) and are critical in the modern defense structure of worldwide military assets. In 2014, the U.S. National Oceanic and Atmospheric Administration (NOAA) publicly stated it had suffered a cyber intrusion, requiring NOAA to seal off much of its data and cut off satellite images for two days.[1] An investigation later found the intrusion compromised certain internal systems, including one system that provides "critical weather satellite data" to the National Weather Service and which serves the "primary forecast centers" for the military.[2] More recently, the U.S. government held a second "hack-a-sat" event in 2021 allowing cyber teams to compete in defending a satellite and attacking an opponent's system in an effort to improve cyber vulnerabilities.[3]

The world today relies heavily on both space and cyber developments. These critical domains utilize evolving technologies and international partnerships to further the national security interests of cooperating nations. Yet, international and domestic laws governing space and cyber assets rely on separate and distinct legal frameworks, effectively creating legal silos within each domain. Regardless, application of some key provisions of the foundational space treaties overlap with cyber operations and should be applied in the cyber threat context – namely, the principles of state responsibility and liability in space.[4] These provisions provide the framework to determine state liability under specific circumstances.

Unique to the space domain, the Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, or Outer Space Treaty, requires states to

1 Mary Pat Flaherty et al., *Chinese Hack U.S. Weather Systems, Satellite Network*, Washington Post (Nov. 12, 2014), https://www.washingtonpost.com/local/chinese-hack-us-weather-systems-satellite-network/2014/11/12/bef1206a-68e9-11e4-b053-65cea7903f2e_story.html?noredirect=on&utm_term=.0f0782a8ef20.

2 *Id.*; Nat'l Oceanic & Atmospheric Admin., Successful Cyber Attack Highlights Longstanding Deficiencies in NOAA's IT Security Program 1 n.2 (2016), https://www.oig.doc.gov/OIGPublications/OIG-16-043-A.pdf.

3 Air Force Research Laboratory Public Affairs, *DAF declares Hack-A-Sat Challenge winners, aims to reduce space vulnerability* (Dec. 13, 2021), https://www.spoc.spaceforce.mil/News/Article-Display/Article/2874198/daf-declares-hack-a-sat-challenge-winners-aims-to-reduce-space-vulnerability.

4 Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies, 610 U.N.T.S. 205, *entered into force*, Oct. 10, 1967 [hereinafter Outer Space Treaty]. "State responsibility" throughout this paper refers to responsibility as outlined in Article VI of the Outer Space Treaty.

assume responsibility for "national activities in outer space," a term undefined in the treaty.[5] In accordance with the Outer Space Treaty provisions, states implement their responsibility through licensing regulations, statutes, oversight of launches, on-orbit activity, and other space-related conduct. Related, the liability provision of the Outer Space Treaty and the subsequent convention governing liability of space objects make a state liable even after a state has transferred ownership of a satellite.[6] States created these remedy processes before complex cyber operations and cyber capabilities were contemplated. Nonetheless, while cyber operations conducted against space objects may complicate the legal remedy process, this paper argues it does not preclude application of traditional space law.

While the writers of the Outer Space Treaty in 1967 surely could not have considered cyber operations as part of these provisions, these responsibility and liability provisions today bridge part of the gap between the legal frameworks mentioned previously. This paper argues legal practitioners can and should apply the Outer Space Treaty provisions of responsibility and liability to cyber operations against space objects. States can further bridge the gap between domains by addressing cyber operations in domestic space law. Our analysis applies during peacetime as laws of war overlap with other aspects of international law during periods of crisis. These applications modernize the practice of law in these areas, reflect the technological reliance on space and cyber domains, and more accurately demonstrate the true reality of multi-domain assets.

To look at a legal framework connecting these two technical domains, this paper first establishes the legal parameters of the responsibility and liability provisions of the Outer Space Treaty. In Section 3, the paper then provides an analysis as to why these also apply to cyber operations in space. In Section 4, this paper gives specific scenarios for these issues and implications for practitioners. While these domains and their legal regimes have remained separated, these provisions connect the domains in a way that can more appropriately place responsibility but also emphasizes the need for domestic legislation.

5 *Id*. at Art VI.

6 *Id*. at Art VII; Convention on International Liability for Damage Caused by Space Objects, *opened for signature* Mar. 29, 1972, 961 U.N.T.S. 187 [hereinafter Liability Convention].

2. SETTING THE STAGE – STATE RESPONSIBILITY AND LIABILITY UNDER INTERNATIONAL LAW

> States Parties to the Treaty shall bear international responsibility for national activities in outer space, including the moon and other celestial bodies, whether such activities are carried on by governmental agencies or by non-governmental entities, and for assuring that national activities are carried out in conformity with the provisions set forth in the present Treaty. The activities of non-governmental entities in outer space, including the Moon and other celestial bodies, shall require authorization and continuing supervision by the appropriate State Party to the Treaty….[7]

A. Background

Articles VI and VII of the Outer Space Treaty outline provisions on state responsibility and liability. A state party to the Outer Space Treaty bears responsibility for its "national activities in outer space."[8] The treaty outlines that state parties to the treaty are also responsible for the national activities of non-governmental entities in the space domain. To extrapolate this further for the modern context, we can see how this necessarily includes cyber activities in space. Setting aside for the moment the question of what constitutes "national activities," it is important to note the application of responsibility toward non-governmental entities was debated during the development of the Outer Space Treaty in the 1960s. On the one side, the Soviet Union was opposed to any private activity in what it considered a national or strategic endeavor.[9] On the other side, the United States, true to its roots in capitalism, did not wish to foreclose the ability of private enterprise to be involved in this new domain. What eventually resulted relevant to non-governmental entities was a compromise between the two schools of thought, in what can be described as "private activity with public responsibility."

Article VI of the Outer Space Treaty reads in part:

> The activities of non-governmental entities in outer space, including the Moon and other celestial bodies, shall require authorization and continuing supervision by the appropriate State Party to the Treaty.[11]

A key legal term in Article VI is "shall," creating a duty on the state party to the treaty

7 Outer Space Treaty, *supra* note 4, 610 U.N.T.S. at Art VI.
8 *Id*.
9 *See* Gennady Zhukov & Yuri Kolosov, International Space Law 64-68 (1985) (discussing a state's responsibility under Article VI, specifically a state's responsibility for private actors in space); Frans von der Dunk, *The Origins of Authorisation: Article VI of the Outer Space Treaty and International Space Law, in National Space Legislation in Europe: Issues of Authorisation of Private Space Activities in the Light of Developments in European Space Cooperation* 5 (Frans G. von der Dunk ed., 2011).
10 *Id*.
11 Outer Space Treaty, *supra* note 4, 610 U.N.T.S. at Art VI.

to "authorize and supervise," and implying states should regulate non-governmental activity in the space domain. Also important to note, "responsibility" in this context differs from the term used in other aspects of international law. Relatedly, Article VII addresses the concept of international liability for damage as follows:

> Each State Party to the Treaty that launches or procures the launching of an object into outer space, including the Moon and other celestial bodies, and each State Party from whose territory or facility an object is launched, is internationally liable for damage to another State Party to the Treaty or to its natural or juridical persons by such object or its component parts on the Earth, in air space or in outer space, including the Moon and other celestial bodies.[12]

A key legal issue under Article VII is the potential for one state or group of states to claim compensation for damage caused by another state or group of states. This liability includes private entities that are subsumed under those states: the attribution of such liability, as per Article VII of the Outer Space and Article I(c) of the Liability Convention, to one or more states takes place regardless of any involvement of private entities in the causation of the damage or the manufacture, launch, or operation of the space object concerned. In other words: one state (or a number of states) will carry the international liability for space activities conducted by private companies.[13]

Further, the Liability Convention names four categories of states who are liable: (1) states that launch a space object, (2) states that procure a launch for a space object, (3) states from whose territory an object is launched, and (4) states from whose facility an object is launched.[14] States later agreed to terms of liability and mechanisms for enforcement in the Convention on International Liability for Damage Caused by Space Objects.[15] This convention includes these same four categories and formally refers to these groups as "launching states."[16]

Finally, the Liability Convention divides liability issues into on-orbit and terrestrial damage. Damage on orbit results in fault liability, meaning a party is liable for any damage as a result of its actions.[17] Terrestrial damage bears absolute liability.[18] Nuances and various scenarios of these issues have been and can be further explored.[19] For purposes of general application of cyber activities to these provisions, it is helpful to simply note the distinction.

12 *Id*. at Art. VII.
13 *Id*.
14 *Id*.
15 Liability Convention, *supra* note 6.
16 *Id*. at 189.
17 Liability Convention, *supra* note 6.
18 *Id*.
19 Frans von der Dunk, Handbook of Space Law 53 (2015); Bin Cheng, Studies in International Space Law 606 (1997) (explaining "such a provision is highly innovatory in international law").

B. The Outer Space Treaty Provides Some Legal Remedies Necessary to Address Cyber Challenges Facing States in the Modern Space Age

Two of the greatest space developments in the modern age are: (1) the continual privatization of space and (2) the increasing reliance of communication systems and other key infrastructure on cyber and space-enabled activities or assets. Cyber activities are key pieces to both of these advancements. Although it would certainly seem useful to develop modernized revisions to the Outer Space Treaty (such as incorporating the realities of cyber effects in the space domain), lack of consensus among the major powers on basic definitions of space security concepts, including what a space weapon is, what constitutes an armed attack in outer space, and the application of the right to self-defense likely prevent meaningful work in this area.[20] For this reason alone, a pragmatic and legal approach suggests looking to the existing Outer Space Treaty, customary practice, and domestic legal regimes to address and solve these modern challenges.[21]

C. State Practice and the U.S. Legal Interpretation of State Responsibility in the Space Domain

The United States has fulfilled its responsibility obligations under the Outer Space Treaty by executing the National Space Policy, National Security Decision Directive Number 42 (July 4, 1982, National Space Policy) and the Commercial Space Launch Act of 1984 (Public Law 98-575, enacted October 30, 1984). In subsequent legislation, the United States has amended its original space policies, which are now codified in the United States Code via Title 51, Chapter 509, 513 and the Code of Federal Regulations 14 C.F.R. ch. III, parts 415, 420, 431 and 435. These regulations govern space activities by companies and individuals in the United States.

Many other states have enacted national legislation,[22] supporting an international perspective that the Outer Space Treaty, Article VI, creates a legal obligation for a state to regulate certain activities by non-governmental entities.[23] Commentators suggest practices adopted by the United States have become legally binding interpretations of Article VI rights and duties.[24]

Given the plain language of Article VI, the background, and state practice, responsibility then means states maintain supervision and licensing of certain space activities, whether conducted by governments themselves or private entities. The

20 Rajagopalan, Rajeswari Pillai. "The Outer Space Treaty: Overcoming Space Security Governance," Int'l Institutions and Global Governance Program, February 2021, https://www.cfr.org/report/outer-space-treaty.

21 Additionally, Articles 31 and 32 of the Vienna Convention on the Law of Treaties proscribe, when determining the meaning of a treaty term, reviewing first the plain language of the treaty, references to the same term in subsequent agreements, and subsequent state practice. Vienna Convention on the Law of Treaties, May 23, 1969, 1155 U.N.T.S. 331.

22 For a list of states with national space legislation, see United Nations Office for Outer Space Affairs, National Space Law, https://www.unoosa.org/oosa/en/ourwork/spacelaw/nationalspacelaw/index.html.

23 Bin Cheng, Studies in International Space Law 606 (1997).

24 Listner, Michael J., SpaceNews Op-Ed, 6 June 2017, https://spacenews.com/a-reality-check-on-article-vi-and-private-space-activities/.

liability provisions name specific liable parties. Later examples will show how these provisions intersect when one party causes damage on orbit and how these apply to specific cyber activities.

3. APPLYING THE TREATY TO CYBER ACTS: WHAT ARE "ACTIVITIES IN OUTER SPACE"?

Although Article VI shows states must maintain responsibility for certain activities, and states have shown they do so through regulation, there is no definition for what those activities are. Further, the prevalence and easy access to cyber operations make liability and responsibility questions more complex. While the responsibility provisions do not specifically name types of operations, a plain reading of the rule and state practice shows what types of operations are included and that certain cyber acts fall squarely under the responsibility provisions.

A. What Constitutes "Activities in Outer Space"?

Whether a state is responsible for certain cyber acts in or affecting the space domain hinges on the application of Article VI. Article VI assigns responsibility to states for "national activities in outer space" and requires non-governmental entities to receive "authorization and continuing supervision" for their activities in outer space.[25] Two aspects of this article require some examination. First, it does not specify what groups of people are covered by the "national activities." Second, both "national activities" and "activities of non-governmental entities in outer space" require an interpretation of what kind of activities are intended.

Article VI does not state who may conduct national activities, other than states and non-governmental entities. Rather, it simply says "national activities." Three schools of thought have arisen suggesting these can be one or more of the following groups of people: a state's nationals, those covered by the Liability Convention, or those over whom a state has general jurisdiction.[26] At a minimum, it is clear that the provision intends to cover both governments and non-governmental organizations. Several nuances to this could be explored, but it is sufficient to say there are conceivable situations where groups of people could fall under any of these categories.

While the groups of people covered by this provision fit into three schools of thought, the types of activities meant by the treaty remain less clear. The term "activities" suggests physical actions conducted in, from, or through outer space. A satellite operator conducts a maneuver by sending signals to the satellite, which changes a

25 Outer Space Treaty, *supra* note 4, 610 U.N.T.S. at Art VI.

26 Bin Cheng, Studies in International Space Law 238 (1997) (explaining these three schools of thought). *See also* von der Dunk, *supra* note 9, at 5.

satellite's physical location to correct or change the orbit of a satellite.[27] To maneuver, a satellite fires thrusters to change the magnitude or direction of the satellite's velocity.[28] Additionally, a launch, the physical means of placing an object in outer space, would meet the intent of the plain language of "activities in outer space." If launches and maneuvers do not constitute "activities in outer space," few activities would.

State practice as governed by domestic legislation further supports states' perceptions of "activities in outer space." As noted, the United States regulates several activities through statute, including launch, satellite communications, and remote sensing.[29] Russia specifically defines "space activity" as "any activity connected with direct conducting of work of exploration and use of outer space including the Moon and other celestial bodies."[30] Other space-faring states have adopted varying definitions of "space activities."[31] For example, legislation from Austria[32] and South Korea[33] discuss launch operations. France,[34] Belgium,[35] and the Netherlands[36] have definitions that include the transitory nature of space objects. The regulation of these space activities suggests states intend to maintain responsibility for certain activities in outer space. While the regulations span from generic to specific, many of them include governance of launches, satellite communications, remote sensing, and satellite maneuvers.[37]

Supporting some of these activities as "national activities," the United Nations has commented on satellite communications and remote sensing, asserting that states bear international responsibility for these activities, further supporting the case for Article VI

27 *Satellite Manoeuvres* European Organisation for Meteorological Satellites, https://www.eumetsat.int/website/home/Satellites/LaunchesandOrbits/SatelliteOrbits/Satellitemanoeuvres/index.html (last visited Aug. 30, 2019).

28 *Id*.

29 National and Commercial Space Programs, 51 U.S.C. §§ 101–713 (2018).

30 Federal'nyi Zakon RF o Kosmicheskoy Deyatel'nosti [Law of the Russian Federation "On Space Activities"], Rossiyskaya Gazeta [Ros. Gaz.] Oct 6, 1993.

31 *See* von der Dunk, *supra* note 19, at 188–204, tbl. 3A.1 (diagraming a table of 21 national space statutes, including definitions of space activities and national activities and other details of each of the 21 states' pieces of legislation).

32 Weltraumgesetz [Space Act] Bundesgesetzblatt [BGBl] No. 132/2011, as amended, https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20007598.

33 Space Development Promotion Act, Act. No. 7538, May 31, 2005, art. 2 (S. Kor.).

34 Loi 2008-518 du 3 juin 2008 relative aux opérations spatiales [Law 2008-518 of June 3, 2008, on Space Operations] Journal Officiel de la Republique Française [J.O.] [Official Gazette of France], Jun. 3, 2008, p. 9169.

35 Loi du 17 septembre 2005 relative aux activités de lancement, d'opération de vol ou de guidage d'objets spatiaux [Law of Sept. 17, 2005, on the Activities of Launching, Flight Operations or Guidance of Space Objects], Moniteur Belge [M.B.] [Official Gazette of Belgium], Jan. 15, 2014.

36 Wet van 24 januari 2007 [Space Activities Act of 2007], Stb. 2007, 80.

37 *See* von der Dunk, *supra* note 19, at 188–204, tbl. 3A.1.

application to these types of space activities.[38] The plain language of the treaty, state application of it, and the United Nations resolutions all suggest we can clearly consider many activities to be "national activities in outer space." A minimum list includes launches, remote sensing, satellite communications, and satellite maneuvers. While other activities likely fall under this provision, these do so with high certainty.

B. Cyber Actions Tantamount to Activities in Outer Space Under Article VI

Largely, parties conduct space activities through cyber means. Signals sent to the satellite for collection of data, transmitting communication, or maneuvering a satellite are sent through networks. For example, a company receives a remote sensing license from its national government to take photos of a portion of the Earth. A group of employees from the company use their software to send a signal through its established network to the satellite. The satellite receives that signal and executes the command to take the photos. The company takes these photos, engaging in remote sensing, pursuant to its license, governed by the state in accordance with Article VI of the Outer Space Treaty. These remote sensing actions constitute national activities in outer space, and the national government is responsible under the Outer Space Treaty for these actions. The government in this scenario has properly provided licensing provisions, and it appears the company has followed the provisions.

At a more granular level, the state has governed the company and its space operators, who enter the command using code, which is sent over a network to the satellite. These commands sent through cyber mechanisms are part of the space activities Article VI intends to govern. Put simply, Article VI of the Outer Space Treaty directly applies to the cyber actions that definitively execute these space activities.

Just as impactful is applying Article VI to a third-party cyber actor. If a company and space operator conduct "national activities in outer space" by executing commands over networks, a third-party cyber actor conducting the same activity without a license must also be engaged in these same space activities. For example, a space operator working for a company sends a signal to a communications satellite pursuant to its license regulated by the state. As discussed, these actions constitute "activities in outer space" requiring state supervision by Article VI. A third-party cyber operator interferes with that signal, sending a different signal to the satellite to execute a different communications command. The third-party cyber operator's actions, though

38 G.A. Res. 37/92, Principles Governing the Use by States of Artificial Earth Satellites for International Direct Television Broadcasting, ¶ 8 (Dec. 10, 1982) (stating "States should bear international responsibility for activities in the field of international direct television broadcasting by satellite carried out by them or under their jurisdiction and for the conformity of any such activities with the principles set forth in this document"); G.A. Res. 41/65, annex, Principles Relating to Remote Sensing of the Earth from Space, at 116 (Dec. 3, 1986) (stating "[i]n compliance with article VI of the [Outer Space Treaty], States operating remote sensing satellites shall bear international responsibility for their activities and assure that such activities are conducted in accordance with the provisions of the Treaty and the norms of international law").

nefarious, send a signal to the satellite in the same manner and to execute the same type of communications command that the space operator's lawful signals do. Although not sanctioned by the state, the cyber operator is conducting these same types of activities in outer space. And the same types of activities are regulated by Article VI.

The implications for cyber actions governed by the Outer Space Treaty become pointed when considering liability and responsibility. Overlapping the liability provisions, the broader application of the responsibility provision suggests states remain responsible for some of these actions. For example, if a third party interferes with the company's signals, sending a different signal to the satellite, the state is still responsible for those actions. The Liability Convention's division between fault liability for on-orbit damage and absolute liability for terrestrial damage creates different nuances to the application of intervening cyber acts. Both are examined below.

1) Cyber Acts Causing Damage on Orbit

State A has domestic space legislation governing, inter alia, launches, satellite communications, and orbital positions. A company receives a license from State A to launch and operate a satellite. As the company operates a satellite from State A, State A remains responsible for the launch and satellite activity under the responsibility provision of the Outer Space Treaty. Sometime after launch, a third party hacks the company's network and sends a command to the satellite to conduct a maneuver. The satellite executes the physical maneuver and crashes into a second satellite launched from State B. Under the Liability Convention, State B seeks recovery of damages against State A.

The Liability Convention outlines fault liability for on-orbit damage.[39] In this case, State A, with evidence of the cyber operation, can claim it is not at fault for the damage. State A would argue although State B properly claims the satellite from State A crashed into its satellite, an intruder actually caused the collision. Likely, negotiations between States A and B will consider whether A was at fault for lacking cyber security and if the cyber operation originated in State A. If the intervening cyber activity originates in another state, State A can show the satellite's damage did not come from State A's actions. State A does not even need to attribute the actions to a specific person or state. Rather, fault liability provisions in the Liability Convention show a state remains liable for damage when it is at fault. If State A demonstrably shows the fault lies with an intervening third party, State A can disclaim liability under the Liability Convention.

2) Cyber Acts Causing Liability for Terrestrial Damage

Fault liability exists in on-orbit situations only.[40] The Liability Convention provides

39 Liability Convention, *supra* note 6.
40 *Id.*

absolute liability for terrestrial damage.[41] Should the third-party cyber actor send a command that results in a collision in the terrestrial territory of State B, State A remains absolutely liable without the option to attribute fault elsewhere. If State A can definitively attribute the cyber action, State A could seek compensation through traditional legal means. Additionally, because the cyber actor engaged in "activities in outer space," the cyber actor's state must have supervision of these activities. Accordingly, if State A can attribute the cyber activity to a certain state, State A can recover damages from that state, relying on the responsibility provisions of the Outer Space Treaty to do so.

3) The "Liability Loophole" and Cyber Responsibility[42]

The complexities continue in considering a potential gap between liability and responsibility. A country could remain liable but no longer have responsibility. For example, in a separate scenario State X launching a satellite remains liable for that satellite no matter what happens to the satellite. If State Y buys that satellite and then operates it, State X remains liable and State Y is responsible. This is sometimes known as the "liability loophole."[43] An additional wrinkle to this loophole is a cyber implication. If a cyber actor from State Z then conducts an operation that results in damage, State X remains liable, State Y could be responsible, and State Z could also be responsible based on the requirement to maintain supervision. The damaged party would have several avenues to recoup damages, the easiest of which is State X under the Liability Convention. State X would then seek to recover those damages from the responsible parties.

C. Limitations

There are limitations to applying these space law provisions to cyber activities. First, attribution remains a challenging aspect of any cyber intrusion, just like it does in other cyber situations. Although states need not fully attribute the cyber activity for damage on orbit, they must show the cyber activity originated from a different state. Further, seeking restitution from another responsible state will require attribution to determine which state remains responsible for those cyber activities.

A second limitation is in state practice. States have not applied this framework to cyber activities, but could clearly do so. Since its signing, states have relied on the Liability Convention just once for resolution, and that case did not result in a claims commission under the terms of the Liability Convention.[44] However, with

41 *Id*.

42 Trevor Kehrer, *Closing the Liability Loophole: the Liability Convention and the Future of Conflict in Space* 20 Chicago J. Int'l L. 178.

43 *Id*.

44 In 1978, a Soviet satellite caused damage in Canadian territory. Canada invoked the Liability Convention, and the parties settled the dispute. Because of the settlement, no Claims Commission formed under the Liability Convention. Peter P.C. Haanappel, *Some Observations on the Crash of Cosmos 954*, 6 J. Space L. 147 (1978); Can. Dep't External Affairs, *Canada: Claim against the Union of Soviet Socialist Republics for Damage Caused by Soviet Cosmos 954*, in 18 International Legal Materials 899, 899 (1979).

technological development, application of the law across these domains will not likely remain as novel.

Finally, there remains no enforcement mechanism for responsibility provisions of the Outer Space Treaty. The Liability Convention allows for a claims commission, but any other aspects require state use of traditional diplomatic measures for enforcement.[45] This can dissuade states from relying on these provisions. However, it should do so no more than other aspects of international law that lack enforcement.

4. IMPLICATIONS FOR STATES AND PRACTITIONERS

Two major implications arise when considering application of these provisions to cyber operations. First, states should anticipate intervening cyber acts as they develop domestic space policy. The Article VI requirement for governments to offer continual supervision over and authorize the "activities of non-governmental entities in outer space, including the Moon and other celestial bodies" requires states to govern even the activities of intervening cyber actors.[46] Cyber acts that execute traditional space activities constitute "activities of non-governmental entities in outer space,"[47] as do those of a company conducting those same activities. While the space licensing requirements govern the state and private actors, this provision in Article VI requires states maintain responsibility for intervening third-party actors who may nefariously engage in cyber acts that constitute activities in outer space. Existing domestic laws governing criminal liability can deter and criminalize this type of conduct within states, but certain conduct may be encouraged or sanctioned by a state. As states develop national legislation, they should consider the responsibility implications of the treaty.

Second, space and cyber law practitioners should incorporate into their practice these provisions as cyber operations against satellites arise. The separation of expertise in these domains leaves challenges for practitioners in understanding the application of various laws. However, Article VI of the Outer Space Treaty sets up responsibility provisions needed to bridge this gap. This provision can deflect liability for on-orbit objects and place financial responsibility on a cyber actor. While attribution is helpful, simply identifying a cyber act as an intervening cause is sufficient to deflect fault liability.

45 Liability Convention, *supra* note 6.
46 Outer Space Treaty, *supra* note 4, 610 U.N.T.S. at Art VI.
47 *Id*.

5. CONCLUSION

With the increasing prevalence of the commercialization of space, the era of space dominance, astropolitics, and space governance is upon us, but we must also not forget the central and linked nature of the cyber threat in that domain. Application of these provisions remains understandably complex and novel. However, as practitioners better understand the gap in technologies and the application of this law, they can further develop approaches to potential recourse. While these rules will likely require diplomatic tools or other means to appropriately negotiate, they should provide practitioners helpful provisions where expertise has traditionally been separated between the domains.

The goal of this paper was to introduce legal issues in assessing responsibility and liability where a space asset has been attacked via cyber means. As legal frameworks and expertise in these domains have remained separated, bridging the gap enables legal practitioners to keep moving toward multi-domain expertise. It is our hope that practitioners are now better equipped to keep moving forward on the application of the law and specifically the application of responsibility and liability provisions of the Outer Space Treaty to challenges in the modern context.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

Maritime Hacking Using Land-Based Skills

Michael L. Thomas
United States Air Force Cyber College
Professor of Cyberwarfare Studies
michael.thomas.180@au.af.edu

Abstract: Today's maritime shipping traffic remains remarkably vulnerable to land cyberattacks due to corporate avoidance of expensive upgrades and maintenance, a lack of skill across the maritime shipping industry, critical maritime chokepoints as ambush locations, and an engrained tendency to view the sea as buffering shipping from land-based threats. When the huge container ship, Ever Given, ran aground in the Suez Canal in March 2021, cyberattacks on the navigation were suspected. Though suspicions were unproven, the effects of the weeklong blockage were global, creating a security warning for national defense professionals. The lesson from the incident is clear – the maritime industry is full of chokepoints that can be economically damaging if exploited. The event also offered a widely covered demonstration event for adversaries seeking future leverage in disruption campaigns. Another demonstration occurred when the winning team at August's 2021 DefCon "Hack the Sea" event hacked replica maritime control systems without the knowledge of ship operational technology. This essay argues that both occurrences are the latest to enhance the recognition of means and opportunity by motivated adversaries. Attackers weaponizing commercial shipping for cybered global disruption need only pick a time and place. A national response is critical to change commercial shipping's cyber defense incentives.

Keywords: *cyber, maritime, computer security, naval cyber*

1. INTRODUCTION

According to the December 2020 National Maritime Cybersecurity Plan [1], the US maritime infrastructure consists of and identifies hundreds of cyber targets, including ports (361), shipyards (124), and tens of thousands of maritime facilities (3500), bridges (20,000), and federal aids to navigation (50,000). The value of maritime trade is understood to be over one-quarter of the total US gross domestic product or over $5.4 trillion. According to the Atlantic Council, worldwide commercial shipping moves close to 80% of global traffic by volume and over 70% of international traffic by value. Despite the slowdown due to COVID, the numbers are expected to increase by over 4% as the economic recovery continues in the next few years [2]. While shore-based infrastructure comprises a considerable segment of the overall infrastructure, what is lacking in many cyber analyses are the cyberattack vectors that are endemic to the platform technology – the vessels themselves. This essay will discuss possible scenarios that might be exploited for nefarious purposes.

A. Scope of the Problem

The sea contains a myriad of users for commercial and military purposes. In general, cargo vessels, bulk carrier container ships, ferries, and oil tankers are used for cargo transportation. Implementing the Global Positioning System (GPS) within the maritime industry revolutionized marine operations by simplifying the architecture of a vessel's bridge navigation systems into an integrated bridge system (IBS). The IBS acts as the central command and control (C2) of a vessel where various digital devices are interconnected for navigation in the open seas. These systems are connected to the other onboard systems of a vessel, including navigation control, propulsion, cargo safety management systems, and administrative and crew email. Additionally, it provides internet access, which leaves the IBS vulnerable and other onboard systems cyber vulnerable. Warnings in various maritime articles have been published frequently over the years [3]. The techniques described often provide adequate details to allow the transfer of techniques from land-based hackers to the maritime environment. The techniques are simple enough – penetrate the platform through onboard navigation and proceed horizontally across other onboard networks to gain control of critical systems such as steering and throttle [4].

In August 2021, hackers did this, and without foreknowledge of the onboard systems, they were able to hack before initiating the penetration. They quickly penetrated the navigation interface and gained control of both the steering systems and the throttle. The result was not a disastrous grounding or a ransom demand but instead, the highly prized "Black Badge" from the Maritime Hacking village of the DEFCON annual cybersecurity conference, held in Las Vegas. DEFCON's "Hack the Sea" hacking challenge has been held over several years with similar results. Year after year,

teams of three to five hackers acquire hands-on experience of hacking real maritime hardware in a controlled environment using the "Grace" maritime cybersecurity testbed provided by Fathom5. The simulated maritime bridge setup provides an accurate facsimile of equipment typically [3] used onboard ocean-going vessels, allowing hacking teams to attack vessels at sea. Using realistic components and protocols, multiple teams of hackers have consistently penetrated different maritime subsystems, including navigation, fire main, and hydraulic steering systems, year after year at various DEFCON events.

2. MODELING THE WORST CASES

While the 2021 challenge and earlier competitions required hackers to connect to propulsion, steering, and navigation systems through a physical wire from their laptops, beginning in 2022, the hope is to provide a wireless environment for the same targets. Notably, the 2021 competition again demonstrated that hacking skills from land-based systems and environments are easily transposable to a maritime environment. The winning team, "The Edmund Fitzgerald," [5] had no experience in the Fathom5 simulator or in maritime hacking in general [6]. A skilled hacking team can typically take less than 14 hours to penetrate system safeguards and remotely compromise both steering and throttle controls. While the scenario used at DefCon required connecting to the simulator, remote access hacking is also feasible, demonstrated in February 2017 [7], when hackers took control of a German-owned container ship sailing from Cyprus to Djibouti. The hackers reportedly compromised both steering and maneuver controls. Control of the steering was only regained by the ship's crew when an information technology team boarded to remediate the cyber breach. Segregation of a vessel's internet protocol networks and serial networks prevents this [8].

The same testbed was used in November 2021 by the US Navy in an evaluation by hackers in the "HACKtheMACHINE" competition to assess the vulnerabilities in the US Navy's unmanned vessel named Sea Hunter, designed by DARPA [9]. Sea Hunter uses both radar and the international ship tracking program known as the Automatic Identification System (AIS) to find its bearings and avoid other vessels. AIS has known vulnerabilities that must be analyzed and resolved before it and autonomous commercial vessels are put into wide use. The Navy's stated goal for such vessels is to have unmanned flotillas operating in the western Pacific by 2030 [9]. Commercial shipping companies are also committed to using unmanned vessels in the commercial maritime sector as well. Given the interest, research, and resources of both the military and commercial maritime industries, cybersecurity is an unstated but obvious requirement before the deployment of such vessels can begin.

As we proceed, it is necessary to imagine possible scenarios that could be utilized by attackers.

A. Scenario 1 – The Maritime Ambush (Background)

The efficient operation of the maritime transportation system (MTS) is under scrutiny by both cyber attackers and defenders. The MTS is at risk more today than ever before. In 2021, cyberattacks targeting the MTS increased by over 400% over a few months [10]. The vast majority of the world's civilian and military maritime traffic passes through a handful of strategic narrow waterways known as "maritime chokepoints." These sea lanes have always been prey to weather, pirates, and maritime accidents. Now added to these traditional perils are maritime cyberattacks – whether motivated by piracy, ransom, malicious disruption, or as part of more significant geopolitical conflicts. Critical shipments get delayed by weeks when an ocean-going vessel is delayed because of hacking. While the Ever Given is one of the largest currently sailing, there are plans to increase the size of such vessels, which will require a further widening of chokepoints [11] like the Suez and Panama Canals and make such targets even more tempting. Container ships as large as 25,000-ton equivalent units (TEUs) are already in the planning stages in shipyards in China and South Korea.

The Suez Canal is one of the more tempting cyber locations simply due to the amount and expected speed of traffic through its one- and two-lane channels – 30% of the world's shipping container volume carrying 12% of global trade passes through the canal [12]. Ships [13] can cut 12 days off a three-week [14] trip from India to Italy by transiting the canal. The 200+ meter wide canal is known to be challenging even at modest transit speeds [15]. For vessels the size of the Ever Given, the 120-mile-long narrow transit offers the opportunity for cyber-induced disruption, particularly if one wants to stall oil and gas deliveries transiting from the Middle East to Europe. A blocked canal means companies must take the alternative route around the Cape of Good Hope, adding 10 to 12 days transit time [16], fuel costs, and security costs [17]. According to a 2006 RAND study [18], the closing of the Malacca Strait increases transit time by only an additional 3–4 days. Disruption caused by closing the Suez Canal is more significant than would be caused by closing the Malacca Straits.

With the grounding of the Ever Given on March 23, 2021, the world was reintroduced to the issue of maritime chokepoints [19]. This is an old naval warfare technique last used by Russia to block the escape route of a part of the Ukrainian Navy when the Russian military annexed Crimea in 2014 [20]. The use of a blocking ship is an effective tactic in naval warfare. The Ever Given blocked the Suez Canal for six days [21]. The Ever Given grounding was not a cyber event, but its grounding demonstrates the negative impact on global trade when a ship blocks it. BBC reported [22] that fears that the blockage would tie up shipments of crude oil resulted in oil prices rising by

4% on world markets. Launched in 2018, the Ever Given remains one of the largest container ships in the world. Built and owned by a Japanese firm, operated and leased by a Taiwanese company, it sails under the Panamanian flag. Similar ships carry an increasing percentage of global trade, and the 2015 addition of a second channel to the Suez Canal was undertaken partly to accommodate [23] them. The widening deepened the main waterway and provided ships with a 35km channel parallel to it. The current main channel is wide enough to accommodate large vessels like the Ever Given, but navigation clearance on either side of both channels is severely limited. Traversing too quickly or misunderstanding wind effects on massive vessels came from human error in this case. The internet protocol (IP) networks used for steering and navigation are often not effectively segregated for cybersecurity [24]. They are connected to the serial bus networks that make up the supervisory control and data acquisition (SCADA) systems critical to onboard operations. The blockage resulting from the grounding of the Ever Given demonstrated to competent cyber terrorists or nation-state adversaries the potential for disruption. The Weiss control system incident database [24] includes more than 30 maritime SCADA cyber incidents. There have been multiple incidents of hacking Global Positioning Systems (GPS) by Russia, China, Iran, and others that have affected ships [25]–[27]. Basic electricity for operating canal locks, such as those in the Panama Canal, offer disruption targets to hackers willing to attack critical infrastructure [28]. The 900 km long Malacca Strait carries 40% of global maritime trade, including a quarter of the globe's seaborne oil supplies [29] and 80% of the Middle East's hydrocarbon supplies to China. Traffic congestion is a challenge, with more than 100,000 vessels transiting the waterway every year, mainly where the strait narrows to just 2.7 kilometers wide [30] in the waters off Singapore. These chokepoints also provide the opportunity, both from the shore and via remote access, for hackers to track targeted ships, owners' fleets, crew, content, origin, destination nationalities, or to select targets. The risks are aggravated as vessels and systems increasingly rely on automation. Fully autonomous ships [31] are a stated goal of the industry [32] and the US Navy [33]. Such systems must consider proper cybersecurity.

Discussion of possible scenarios that attackers could utilize in this environment is necessary.

B. Scenario 1 Continued – The Maritime Ambush (Maritime Choke Points Make Natural Kill Zones)

Imagine a US Navy carrier strike group transiting the Suez Canal. In the US Navy carrier air operations, a carrier strike group usually consists of 1 aircraft carrier, 1 guided missile cruiser (for air defense), 2 Light Airborne Multipurpose Systems (LAMPS) capable warships (focusing on antisubmarine and surface warfare), and 1–2 destroyers or frigates. This package is designed for military action in deep water

open ocean. As it transits the canal, a blocking ship could be sunk ahead, and an attack from a surface-based small seacraft loaded with explosives could then be deployed. One need only recall the damage done to the USS Cole to know that a small craft laden with enough explosives could create severe damage. If the carrier-launched aircraft to assist in defense of the battle group, these could be nullified by personnel stationed ashore with ground-to-air missiles of a stinger capability. While the engagement would not be expected to last long, the potential damage to the flotilla and accompanying casualties could be severe.

1) Ships and Cyber Security Need an Introduction

In June 2018, security researchers at Pen Test Partners [34] found vulnerabilities in the electronic chart and mapping display and information systems (ECDIS) commonly used on cargo and container ships. When these chart systems are linked to GPS-enabled autopilots [35], this gives hackers the ability to access the critical ops technology Operational Technology (OT) of the vessel. Hackers can remotely operate the ship's steering, navigation, and ballast pumps if networks are not segregated. ECDIS is often linked directly to the autopilot on vessels, causing the ship to follow the programmed course. Hackers can redirect the vessel's course by inserting false information messages via Satcom [36] to mislead navigational decisions [37]. Many satellite terminals on vessels are available on the commercial internet with default credentials and can be remotely hacked [36]. Multiple other paths can also prove useful vectors in the cyberattack of a ship. In 2018, research showed that the ECDIS systems on some vessels were still using relic operating systems [8] with many known major vulnerabilities, such as Windows NT, often because these are expensive to upgrade. Even when malicious control is discovered, it can be tough to regain control promptly.

Commercial ship networks [24] often have flat network architectures that are originally unsegmented and without firewalls or other cybersecurity means as part of the architecture. Once inside the networks, it is not difficult to traverse all the systems of the entire ship. Researchers often identify other vulnerabilities in computer security forums – for example, using the vessel's Satcom terminal as a point of entry [38]. The Satcom terminal opens the system to attackers, and malware can replace the poorly secured firmware or simply revert to an even less secure earlier version and then alter the applications running on the terminal. Access, whether via ECDIS, the Satcom terminal, or any other outward-facing comms, provides a means to take control of critical ship systems covertly and use the massive bulk of the vessel for any purpose the hacker chooses.

Some speculated that the Ever Given accident was a cyber incident from the beginning. This theory was shown to be incorrect when the voyage data recorder was examined.

In rebuttal, as experienced cyber control systems expert Joe Weiss [24] points out, the potential for cyber disruption still remains. Despite the Ever Given being only three years old and the latest marine electronics likely installed for control and navigation, this will not resolve the vulnerabilities discussed earlier. The recent DefCon exercise is not a stand-alone example of success in simulated maritime hacking – it has been repeated over several years. Concurrent with the Ever Given grounding, a team of Ph.D. students that competed in an earlier naval "HACKtheMACHINE" [39] exercise (using the same "Grace" maritime system as the earlier DefCons) attempted to determine if hackers could successfully attack maritime systems remotely through a cloud network. The team succeeded [38], "hacking and crashing the [fictional ship's] cybersecurity monitoring system." While a wireless attack was not repeated at the August 2021 DefCon, it shows attacks through the cloud are possible.

2) Shipping as a Cyber Weapon

Hackers will not ignore the opportunities presented by poor cybersecurity. The transportation vector has already been exploited in non-cyber scenarios. One only has to think back to the September 11 attacks on the twin towers in New York City to realize that this is a viable attack vector. A cyber campaign will provide a good enough return on investment in economic or political benefits to make it attractive, and potentially lucrative. US adversaries such as Russia, China, and Iran learn from these exploits and integrate them in more extensive cyber-enabled campaigns. Russia, for example, spoofed ship GPS systems [40] thousands of times between 2016 and 2019, affecting around 1,300 commercial ships. In 2017, DPRK navigation jamming was the culprit behind the return of hundreds of Republic of Korea (ROK) fishing vessels, and such cyberattacks led to the devastating NotPetya attacks that crippled the Maersk shipping line the same year [41]. In July 2021, Britain's Sky News reported having acquired documents said to have come from an Iranian offensive cyber unit [42] called Shahid Kaveh, which is part of the Islamic Revolutionary Guard Corps (IRGC) cyber command. They present specific research on how to sink a cargo ship using cyberattacks and include details on the Satcom systems used in the international shipping industry. They are likely creating a target set for use at some later date.

C. Scenario 2 – Destruction of a Major Urban Area Using Liquid Natrual Gas (LNG) Infrastructure

Regarding the earlier mentioned German loss of control of a container vessel for 10 hours, imagine such a compromise of an LNG tanker in a large urban area. There are 9 LNG terminals in the US, and Boston is the only major city in the US where LNG infrastructure is located near such a large urban area. The location of this terminal makes it an ideal candidate for a terrorist attack. Weekly, LNG tankers pass within half a mile of the crowded Boston waterfront, past the end of the Logan International Airport runway, and under a busy bridge [43]. In a 2004 Sandia Lab study sponsored

for the Department of Energy (DOE) [44], the lab postulated and modeled a worst-case scenario, and "indicates that a successful attack on a tanker – via methods such as internal sabotage, a rocket-propelled grenade, a kamikaze flight, or a USS Cole style suicide boat ramming – would create a profound security threat for a city like Boston." Such an attack vector would not necessarily require anything more than a few hijacked tugboats laden with high explosives. Such tugs are typically used to control a vessel within a narrow strait and are used in Boston harbor for this very purpose.

The DOE has estimated that if there were an LNG tanker incident in the harbor, up to 80,000 people would die within the first 20 seconds, and upwards of another 500,000 would suffer severe burns in a 2.5-mile radius of the harbor within the first 8 minutes of such an event. The loss of control of a tanker, onshore LNG storage tanks, and an LNG pipeline all pose significant risks for city residents if compromised [45].

Al-Qaeda's attack on the USS Cole on October 12, 2000, in the harbor in Aden, Yemen, changed the calculus. A small inflatable boat loaded with explosives was used to blow a 40×60-foot hole in the side of the armored ship. They inflicted heavy damage, killed 17 USN personnel, and injured 36 others. Shortly before this, a small boat laden with explosives attacked the French Limburg tanker at Ash Shihr, Yemen. Both the hulls of the double-hulled ship were penetrated, and damage, according to the captain, extended "[s]even or eight meters into the cargo hold, which was filled with crude oil."

LNG tankers are also double-hulled. The major difference to oil tankers is that the LNG cargo is contained in a series of 3–5 tanks that maintain the cargo at -260°F at normal atmospheric pressure. While safe to transport LNG, tankers are not armored and designed to withstand deliberate attack. The tanks are covered by insulating foam that is both flammable and fragile. Designed and constructed to meet the international "Gas Tanker Code" they must meet the US Coast Guard (USCG) "Type IIG" standard of subdivision, damage stability, and cargo tank location. Neither standard addresses a design to minimize the consequences of an intentional terrorist attack on an LNG tanker. Additionally, the USCG maintains an "exclusion zone" around LNG tankers during port operations. Monitoring and compliance require only that the owner/operators of LNG tankers and port facilities maintain responsibility for security training and procedures, and a current security certification by the owner/operator be submitted to the USCG.

The typical LNG tanker often travels at speeds above 20 knots on the open sea. In port, they must maintain much lower speeds and often rely on tugs to maneuver. In contrast to small outboard motor-driven boats that can achieve speeds over 40 knots.

and turn within their own lengths, LNG tankers are seagoing "sitting ducks." If a group of terrorists has the tactical objective of a coordinated attack on an LNG tanker or port facility, the initial phase of the operation would be the acquisition of a suitable small vessel. This is trivial and might even be a small hijacked or pirated tug.

After planning, the next step is to acquire appropriate explosives. Obtaining the materials is not difficult. For the attack on the Murrah Building in Oklahoma City, a devastating bomb was easily concocted from common fuel oil and ammonium nitrate fertilizer, which can be obtained at any farm supply outlet in large quantities. Other explosives as well as detonators are also easy to obtain, particularly for today's well-funded and geographically dispersed terrorist organizations.

Lastly, to complete planning an attack on an LNG tanker, it is necessary for the terrorist group to obtain precise intelligence on the design, current location, and itinerary of a target LNG tanker. Without consideration for potential misuse, the internet provides an ideal tool: the Vessel Tracking Information System (VTIS.)

VTIS is a real-time traffic information system utilized by the MTS in many of the world's shipping ports. VTIS systems rely on existing GPS receivers and communications to instantly transmit position, course, and related information. The systems also provide a visual display of all large ships in the harbor and compute collision avoidance tracks. With real-time information on the heading, position, and speed of an LNG tanker, together with information on most of the other vessels in the port, executing a Cole-style attack becomes relatively simple.

If a terrorist decided to use an LNG tanker as a weapon, how bad would it be? That has been considered as far back as 2004. Currently, the US military's largest non-nuclear weapon is the "daisy cutter" BLU-96. It disperses 2,000 lbs. of a flammable hydrocarbon, has a blast zone of over 500 feet in radius, and consumes all available O2 within that zone and for some distance beyond. Release of the liquid can form a pool on the surface of water, and when this is ignited, start a "pool fire." This fire would create a large amount of radiant heat. Some estimates suggest the radius of such a fire could extend out to as far as 1.25 miles [46]. Compare the BLU-96 with the 130,000 cubic meters of LNG contained in a typical tanker: 3068532 MBTUs (million British thermal units) of energy. *This is the equivalent of 775 kilotons of TNT*. The Hiroshima bomb had a yield of 15 kilotons of TNT. This illustrates the extreme danger to a crowded urban area like Boston [47]. Terrorists do not need to acquire nuclear materials to achieve nuclear device results.

D. Future Attack Vectors

The routine hacking of vessels using satellites is imminent. The current Global Navigation Satellite System (GNSS) constellation includes the American run GPS, the Russian GLONASS, the European Union's GALILEO, China's BeiDou, Japan's Quasi-Zenith Satellite System (QZSS), and India's Regional Navigation Satellite System (IRNSS), (with an operational name of *NavIC*) system [48]. Each nation's ships tend to use their national system, and no nation's commercial vessels are as secure as they currently need to be. They also lag in securing shipboard systems in the near and medium term. There is talk of using older radio wave technology as a more secure alternative to Satcom systems, but the discussions are only at the initial stages. In 2018, President Trump signed the National Timing Resilience and Security Act into law as part of the Frank LoBiondo Coast Guard Authorization Act. This legislation mandates the Secretary of Transportation to establish a land-based timing system to provide a backup for GPS, to "ensure the availability of uncorrupted and nondegraded timing signals for military and civilian users." It is questionable how rapidly alternatives such as eLORAN [49] will spread. As one researcher states, [50] "[Electronic charting] systems pretty much never have antivirus." The antivirus (AV) industry [51] that protects land-based computers in the US and Europe was started over 30 years ago, but countless huge ships launched during that time with complex computer architectures contain only basic AV protection. Keeping machines on land updated with current malware definitions is problematic. Doing so with moving targets on a deadline to move cargo is exponentially more complicated for a number of reasons, the lack of trained merchant seamen able to perform the task being only one of the complications.

Military and civilian shipping worldwide plan on free transit through the Suez Canal and other chokepoints. Iranian intelligence has collected maps, means, and incentives to use maritime cyber weaknesses in Iranian campaigns. In the 1990s, the bin al-Qaeda group experimented with a plethora of attempted attacks using public transit, notably in Paris. Six years later, al-Qaeda used airliners against the Twin Towers in New York City on September 11. The technical means to exploit cyber insecurities are well distributed across land-based hackers with no prior maritime experience. The motivation varies as much as the adversary, ranging from the ransomware criminal to the "just because I can" opportunist to the state adversary or its proxies.

3. CONCLUSIONS

The gauntlet has been thrown down for westernized democracies to ignore or pick up. US shipping [52] has yet to begin to address the cybersecurity issues and international disruption from hacking modern container ships. Both President Biden's

administration and Republican definitions of infrastructure miss a key dimension of modern connectivity: what happens onshore often begins at sea. As on land, cybersecurity needs to be practiced by users in all sectors of maritime – not just IT experts. What is a problem on land is a problem at sea, and maritime problems are often more complex to solve. Competent maritime IT experts also need a degree of domain expertise – something more challenging in the maritime sector. Ninety percent of the world's trade travels by sea [53] and 40 million US jobs [54] depend on trade. In military strategic thought, the triad of means, opportunity, and motivation only lack the final "when." The adversary gets the vote as to this when. An uptick in cyberattacks targeting maritime targets includes varieties of attacks familiar to land-based targets, including phishing, ransomware, and various forms of malware. Combined with traditional cyber threats targeting information technology (IT) systems, reports of attacks on operational technology (OT) and platform technology (PT), such as ships and ports, increased an estimated 900 percent in the three years to 2020 [2].

Strategic national security actions must include dramatic changes in commercial shipping incentives that ensure cyber defenses for ships. The threats to maritime traffic exist and cannot be ignored. They transcend merely assuring infrastructure for the US Navy. Serious national security responses must include the carrot and the stick. At a minimum, requiring proof of cybersecurity for container and other commercial vessels that enter US national waters, and increased federal financial support for cybersecurity for the ports, shipping, and shipbuilders that serve the needs of the US maritime industry, is necessary.

The US maritime industry should adopt a "system of systems" approach to maritime cybersecurity and extend the 2020 National Maritime Cybersecurity Plan (NMCP) and proposed legislation in the current negotiations behind the SHIPYARD Act, beyond ports alone to also include container ships as a necessary first step.

First, it should be recognized that ships are systemically insecure. This does not mean to imply that a one size fits all approach to security standards would solve such a vessel-specific problem. The global fleet is neither homogeneous nor monolithic. Transnational organizations must continue to implement the existing National Institute of Standards and Technology (NIST) Cybersecurity Framework within the MTS. One complication to shipping is the fact that the fleet being in motion makes achieving a baseline difficult and implementing frequent updates to systems at sea more than trivially problematic.

Second, there is a current lack of transparency, situational awareness, and collaboration in the MTS that defines the adversary's ability to define emerging threats and maximize risk on both a sector-wide and subsystem-specific basis. It is necessary to

consider that the MTS expand and clarify the necessary protocols and programs to streamline, aggregate, and make transparent and incentivize information exchange and vulnerability disclosure. It must be more than mere information exchange.

Third, education and training requirements abound across the public and private sectors, and more MTS-specific cybersecurity education is needed. Proper training and education on maritime cyber threats and cyber best practices can make a difference when it comes to risk mitigation.

Lastly, ships are the heart of the MTS, but the variety defined by the nature of missions and diversity of systems sets them up for failure and makes the task of the widespread adoption of cybersecurity best practices almost impossible. This aside, addressing the vulnerabilities in ports and land-based cranes is not sufficient. It is impossible to increase the security of the broader MTS without also addressing shipboard cybersecurity.

The Atlantic Council's Report "Raising the Colors" [2] offers recommendations for shore-based sections of the MTS infrastructure but offers little to solve the very real issues of the ships themselves. Military ships are the responsibility of the Navy, and this is being addressed. In the civilian sector, new policies should require proof of implemented procedures and provide funding for cybersecurity upgrades in all container vessels delivering cargo to US ports. This is a response that ought to be in alignment with other seafaring nations. United with our traditional allies, the US government can uniquely influence what is considered normal but is inadequate in the construction, operation, and insurance of the global maritime fleet. The US and its allies are majority stakeholders in the global maritime trade environment system. It is the same system that the US's major adversary, China, intends to dominate in coming decades with vessels, export volumes, ports, political coercion, and military saber rattling. Either the US directly addresses these issues with stakeholders, or it will spend much more in lives when adversaries attack at the place and time of their choosing. Both DefCon "Hack the Sea" and the US Navy "HACKtheMACHINE" exercises have demonstrated how to gain control of a vessel via cyber means and weaponize the vessel. This is going to occur sooner or later.

REFERENCES

[1] United States Department of Homeland Security, "2020 National Maritime Cyber Security Plan," Washington DC, USA, White House Office. [Online]. Available: https://homeport.uscg.mil/Lists/Content/Attachments/65433/National%20Maritime%20Cybersecurity%20Plan%20to%20the%20National%20Strategy%20for%20Maritime%20Security%20Dec%202020.pdf

[2] W. Loomis, V. V. Singh, G. C. Kessler, and X. Bellekens, “Raising the colors: Signaling for cooperation on maritime cybersecurity,” Scowcroft Center for Strategy and Security, Idaho Falls: Atlantic Council Cyber Statecraft Initiative, 2021. [Online]. Available: https://www.atlanticcouncil.org/wp-content/uploads/2021/10/Cyber-Maritime-Final-Report.pdf
[3] E. Montalbano, “Container ships easy to hack, track, send off course and even sink, security experts say,” Jun. 5, 2018. [Online]. Available: https://securityledger.com/2018/06/container-ships-easy-to-hack-track-send-off-course-and-even-sink-security-experts-say/
[4] D. Storm, “Hack in the Box: Researchers attack ship tracking systems for fun and profit,” Computer World, Oct. 21, 2013. [Online]. Available: https://www.computerworld.com/article/2475227/hack-in-the-box--researchers-attack-ship-tracking-systems-for-fun-and-profit.html
[5] DEFCON 2021, “DEFCON 29 Score Rankings,” Aug. 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.sea-tf.com/
[6] A. Chaveriat, *Hack a Boat [SEA-TF: Maritime Hacking] Contest | DEF CON 29*. (2021). [Online Video]. Available: https://www.youtube.com/watch?v=Hi3Nra4QrRk
[7] T. Blake, “Hackers took ‘full control’ of container ship’s navigation systems for 10 hours,” Nov. 22, 2017. Accessed: Nov. 20, 2021. [Online]. Available: https://www.asket.co.uk/post/2017/11/26/hackers-took-full-control-of-container-ships-navigation-systems-for-10-hours-asketoperati
[8] K. Munro, “Hacking Serial Networks on Ships,” Jun. 25, 2018. Accessed: Nov. 20, 2021. [Online]. Available: https://www.pentestpartners.com/security-blog/hacking-serial-networks-on-ships/
[9] J. Vincent, “The US Navy’s new autonomous warship is called the Sea Hunter,” Apr. 8, 2016. [Online]. Available: https://www.theverge.com/2016/4/8/11391840/us-navy-autonomous-ship-sea-hunter-christened
[10] Security Magazine, “Maritime Industry Sees 400% Increase in Attempted Cyberattacks Since February 2020,” Jun. 8, 2020. Accessed: Nov. 20, 2021. [Online]. Available: https://www.securitymagazine.com/articles/92541-maritime-industry-sees-400-increase-in-attempted-cyberattacks-since-february-2020
[11] D. Fickling, “Giant Next-Gen Container Ships Will Make Ever Given Look Like Toy,” Mar. 29, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.bloombergquint.com/opinion/despite-the-ever-given-getting-stuck-in-the-suez-canal-ships-will-get-bigger
[12] A. Veiga, “Suez Canal blockage adds to pressure points in global trade,” Associated Press, Mar. 29, 2021. Accessed: Nov. 20, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://apnews.com/article/europe-global-trade-egypt-coronavirus-pandemic-suez-canal-166bc8f21e9705f2921a67ef2dea176c
[13] F. Bahtić, “World’s largest containership makes its first crossing through Suez Canal,” Aug. 31, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.offshore-energy.biz/worlds-largest-containership-makes-its-first-crossing-through-suez-canal/
[14] R. Picheta, “Why the Suez Canal is so important – and why its blockage could be so damaging,” Mar. 26, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.cnn.com/2021/03/26/africa/suez-canal-importance-explainer-scli-intl/index.html
[15] S. McNeice, “‘All Hell Breaks Loose’ – What It’s Like to Navigate Through the Suez Canal,” Newstalk, Mar. 29, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.newstalk.com/news/all-hell-breaks-loose-what-its-like-to-navigate-through-the-suez-canal-1172045
[16] D. Bernardy, “Approximately how much travel time was saved by the opening of the Suez Canal in 1869?” April 5, 2019. Accessed: Nov. 20, 2021. [Online]. Available: https://history.stackexchange.com/questions/51958/approximately-how-much-travel-time-was-saved-by-the-opening-of-the-suez-canal-in
[17] G. Topham, “How the Suez canal blockage can seriously dent world trade,” Mar. 26, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.theguardian.com/business/2021/mar/26/how-the-suez-canal-blockage-can-seriously-dent-world-trade
[18] M. D. Greenberg, P. Chalk, H. H. Willis, I. Khilko, and D. S. Ortiz, “Maritime Terrorism: Risk and Liability,” Rand Corp, 2006. Accessed: Nov. 20, 2021. [Online]. Available: https://www.rand.org/content/dam/rand/pubs/monographs/2006/RAND_MG520.pdf
[19] MI News Network, “What are Maritime Chokepoints?” Oct. 21, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.marineinsight.com/marine-navigation/what-are-maritime-chokepoints/
[20] A. Kermenchikli and M. Lvovski, “Второй ‘раздел’ Черноморского флота: Украина сохранила 10 кораблей из 61” [The second ‘section’ of the Black Sea Fleet: Ukraine retained 10 ships out of 61], Mar. 26, 2014. Accessed: Nov. 20, 2021. [Online]. Available: https://www.segodnya.ua/regions/krym/vtoroy-razdel-chernomorskogo-flota-ukraina-sohranila-10-korabley-iz-61-505582.html
[21] M Henley and J. Safi, “How a full moon and a ‘huge lever’ helped free Ever Given from Suez canal,” Mar. 30, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.theguardian.com/world/2021/mar/30/powerful-tugs-and-an-ebbing-tide-how-the-ever-given-was-freed
[22] T. Leggett, “Egypt’s Suez Canal blocked by huge container ship,” Mar. 24, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.bbc.com/news/world-middle-east-56505413

[23] BBC, “Egypt launches Suez Canal expansion,” Aug. 6, 2015. Accessed: Nov. 20, 2021. [Online]. Available: https://www.bbc.com/news/world-middle-east-33800076
[24] J. Weiss, “Was the Ever Given hacked in the Suez Canal?” Apr. 13, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.controlglobal.com/blogs/unfettered/was-the-ever-given-hacked-in-the-suez-canal/
[25] J. Edwards, “The Russians are screwing with the GPS system to send bogus navigation data to thousands of ships,” Apr. 14, 2019. Accessed: Nov. 20, 2021. [Online]. Available: https://www.businessinsider.com/gnss-hacking-spoofing-jamming-russians-screwing-with-gps-2019-4
[26] J. Trevithick, “New Type Of GPS Spoofing Attack In China Creates ‘Crop Circles’ Of False Location Data,” Nov. 18, 2019. Accessed: Nov. 20, 2021. [Online]. Available: https://www.thedrive.com/the-war-zone/31092/new-type-of-gps-spoofing-attack-in-china-creates-crop-circles-of-false-location-data
[27] M. Kumar, “Iranian engineer hijack U.S. drone by GPS hack [Video Explanation],” Dec. 16, 2011. Accessed: Nov. 20, 2021. [Online]. Available: https://thehackernews.com/2011/12/iranian-engineer-hijack-us-drone-by-gps.html
[28] C. Duffy and Iyengar Rishi, “Hackers have a devastating new target,” Jun. 4, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.cnn.com/2021/06/03/tech/ransomware-cyberattack-jbs-colonial-pipeline/index.html
[29] S. D. Mateus, “‘Worrying’ rise in piracy attacks around Malacca Strait,” Nov. 11, 2014. Accessed: Nov. 20, 2021. [Online]. Available: https://www.dw.com/en/worrying-rise-in-piracy-attacks-around-malacca-strait/a-17780275
[30] N. Martin, “Suez Canal blockage: 4 of the biggest trade chokepoints,” Mar. 27, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.dw.com/en/suez-canal-blockage-4-of-the-biggest-trade-chokepoints/a-57020755
[31] Korea Advanced Institute of Science and Technology, “Autonomous ships for the high seas,” Mar. 30, 2016. Accessed: Nov. 20, 2021. [Online]. Available: https://www.sciencedaily.com/releases/2016/03/160330182854.htm
[32] R. Beighton, “World’s first crewless, zero emissions cargo ship will set sail in Norway,” August 27, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.cnn.com/2021/08/25/world/yara-birkeland-norway-crewless-container-ship-spc-intl/index.html
[33] USNI News, “The US Navy officially received the DARPA Sea Hunter,” Apr. 13, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://military-wiki.com/the-us-navy-officially-received-the-darpa-sea-hunter/
[34] Cyware Hacker News, “Hackers could monitor, hijack, steal and even sink ships by exploiting flaws and poor security,” Jun. 12, 2018. Accessed: Nov. 20, 2021. [Online]. Available: https://cyware.com/news/hackers-could-monitor-hijack-steal-and-even-sink-ships-by-exploiting-flaws-and-poor-security-3b70c9f5
[35] UT News, “Spoofing a Superyacht at Sea,” Jul. 30, 2013. Accessed: Nov. 20, 2021. [Online]. Available: https://news.utexas.edu/2013/07/30/spoofing-a-superyacht-at-sea/#:~:text=Spoofing%20is%20a%20technique%20that%20creates%20false%20civil,the%20ship%E2%80%99s%20command%20room%20could%20identify%20the%20threat
[36] R. Santamarta, “SATCOM Terminals: Hacking by Air, Sea, and Land,” 2014. Accessed: Nov. 20, 2021. [Online]. Available: https://www.blackhat.com/docs/us-14/materials/us-14-Santamarta-SATCOM-Terminals-Hacking-By-Air-Sea-And-Land-WP.pdf
[37] R. Heilweil, “For hackers, space is the final frontier,” Jul. 29, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.vox.com/recode/22598437/spacex-hackers-cyberattack-space-force
[38] T. Carnicelli, “Warfare Center’s Cyber Red Team Notches Another Win in National Hacking Event,” Apr. 26, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.dvidshub.net/news/394772/warfare-centers-cyber-red-team-notches-another-win-national-hacking-event
[39] C. Villarreal, “NSWC Corona Scientists, Engineers Snag ‘Hack the Machine’ Win,” Mar. 29, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.dvidshub.net/news/392535/nswc-corona-scientists-engineers-snag-hack-machine-win
[40] M. Burgess, “To protect Putin, Russia is spoofing GPS signals on a massive scale,” Mar. 27, 2019. Accessed: Nov. 20, 2021. [Online]. Available: https://www.wired.co.uk/article/russia-gps-spoofing
[41] M. McQuade, “The Untold Story of NotPetya, the Most Devastating Cyberattack in History,” Wired, Aug. 22, 2018. Accessed: Nov. 20, 2021. [Online]. Available: https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/
[42] D. Haynes, “Iran’s Secret Cyber Files,” Jul. 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://news.sky.com/story/irans-secret-cyber-files-on-how-cargo-ships-and-petrol-stations-could-be-attacked-12364871

[43] C. Hurst, "The Terrorist Threat to Liquefied Natural Gas: Fact or Fiction?," Institute for the Analysis of Global Security (IAGS), Washington DC, USA, 2008. Available: https://apps.dtic.mil/sti/pdfs/ADA477509.pdf
[44] C. Savage, "Study spells out high toll on city in LNG attack," Dec. 21, 2004. Accessed: Nov. 21, 2021. [Online]. Available: http://archive.boston.com/news/local/articles/2004/12/21/study_spells_out_high_toll_on_city_in_lng_attack/
[45] Neighborhood Boston, "Liquefied Natural Gas (LNG) fire risk," 2018. Accessed: Nov. 21, 2021. [Online]. Available: http://boston.neighborhoodx.com/lists/index?g=118
[46] M. N. Murphy, *Small Boats, Weak States, Dirty Money*. London, UK: Hurst Publishers Ltd., 2020.
[47] L. Husick and S. Gale, "Planning a Sea Borne Terrorist Attack," Mar. 1, 2005. Accessed: Nov. 21, 2021. [Online]. Available: https://www.fpri.org/article/2005/03/planning-a-sea-borne-terroris-attack/
[48] GPS.GOV, "Other Global Navigation Satellite Systems (GNSS)," Oct. 19, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.gps.gov/systems/gnss/
[49] J. Whitney, "Public-private partnership to launch eLORAN technology to back-up and accompany GPS," Aug. 12, 2020. Accessed: Nov. 20, 2021. [Online]. Available: https://www.militaryaerospace.com/rf-analog/article/14181490/eloran-loran-c-gps-gnss
[50] J. Chesaux, "Cyber Attacks at Sea: Blinding Warships," Jul. 2, 2020. Accessed: Nov. 20, 2021. [Online]. Available: https://www.cyberdefensemagazine.com/cyber-attacks-at-sea-blinding-warships/
[51] M. Sahay, "Who Invented the Antivirus? A History of Antivirus Software," Oct. 29, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.thepcinsider.com/who-invented-antivirus-history-timeline-evolution/
[52] A. Klein and B. Jones, "Why maritime infrastructure is about more than the U.S. Navy," May 21, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://www.brookings.edu/blog/fixgov/2021/05/21/why-maritime-infrastructure-is-about-more-than-the-u-s-navy/
[53] E. Kimball, "How securing maritime commerce protects U.S. national security," Apr. 1, 2019. Accessed: Nov. 20, 2021. [Online]. Available: https://www.brookings.edu/blog/order-from-chaos/2019/04/01/how-securing-maritime-commerce-protects-u-s-national-security/
[54] Business Roundtable, "New Study: Trade Supported Over 40 Million American Jobs," 2020. Accessed: Nov. 20, 2021. Accessed: Nov. 20, 2021. [Online]. Available: https://s3.amazonaws.com/brt.org/Trade_and_American_Jobs_2020.pdf

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

A Cryptographic and Key Management Glance at Cybersecurity Challenges of the Future European Railway System

Mikko Kiviharju
Finnish Defence Research Agency
Riihimäki, Finland
mikko.kiviharju@mil.fi

Christina Lassfolk
CCD COE
Tallinn, Estonia
christina.lassfolk@ccdcoe.org

Sanna Rikkonen
Finnish Defence Research Agency
Riihimäki, Finland
sanna.rikkonen@mil.fi

Hannu Kari
National Defence University
Helsinki, Finland
hannu.kari@mil.fi

Abstract: In today's globalized economy, the transport of people and goods is essential. Railways form a critical infrastructure that constitutes one of the backbones of modern society. Furthermore, a shift from air and road transport to electric railway enables a cost-efficient reduction of the negative climate impact caused by transport. Thus, the utilization of existing railway systems must become more efficient. Consequently, the importance, complexity, and smartness of these systems increases, but the systems simultaneously become more vulnerable to attacks.

Within the European Union, there is a quest for a pan-European standardized control system for railway traffic. Old, national, non-interoperable railway communication systems need renewal. The European Rail Traffic Management System (ERTMS) is the European standard for automatic train protection as well as for command and control systems. The next-generation development effort, the Future Railway Mobile Communication System (FRMCS), builds on a 5G service standard. This article focuses on two key aspects of cybersecurity, namely cryptographic and key management challenges in proposed railway communication systems.

Keywords: *cybersecurity, railway, communication, cryptography*

1. INTRODUCTION

The annual growth rate of railroad transportation, in terms of passenger kilometers or cargo tonne-kilometers, has been moderate in comparison with air transportation [1]–[5], and it has a similar rate to road transportation [6], [7]. When comparing carbon dioxide (CO_2) emissions of alternative transport systems, trains outperform airplanes, cars, trucks, and buses [8]. Hence, a shift from air and road transport to electric railway transport is a cost-effective and efficient way to reduce the negative climate impact of transport infrastructures. Naturally, this shift will not be complete. Instead, all three will remain and coexist as a cost-efficient transportation system of the future.

We will see a dramatic increase in the demand for rail transportation for both passengers and cargo. As the total amount of rail lines decreases [9] rather than increases, there is increased pressure to improve the efficiency of the existing rail transportation system. As part of securing operations of critical infrastructure of modern society, this will lead to new security and safety challenges as we must utilize existing railroads more efficiently (i.e., safely, but closer to each other, running more trains on the same tracks).

Similarly, as for road and air transportation, both passengers and goods can be transported at the same time via rail. However, separation of passenger and cargo transportation is common, as it enables more optimized logistics for these segments with differing response time requirements and logistical chains. A comparison of these three transportation infrastructures reveals that the rail system is closer to the air transportation infrastructure in terms of regulation, standards, and the number of organizations involved, whereas the road infrastructure is less controlled and regulated and is clearly an open multi-player environment.

On the roads, cars and trucks of arbitrary age share the same infrastructure. There is loose control of the skills of drivers with driver's licenses and annual inspections of vehicles. In contrast, rail and air traffic are very strictly controlled. Communication protocols, type approval of equipment, and qualified operators are mandatory both in rail and air traffic environments for safe operation. Interoperability of air traffic is mandatory, as it is truly a global business, whereas rail traffic complies with national or continental standards. Especially within the European Union, there is a quest for a Europe-wide standardized rail traffic control system [10], [11].

From the operation's point of view, air and rail traffic are quite similar to each other and are similar in each country. There are a few active operators in these environments, nonetheless much fewer than in road traffic. Air and rail traffic have two main differences. First, even though air space congestion is considerable in some

parts of the globe, it is possible to redirect flights into different routes, thus allowing aircraft to avoid collisions and let the faster aircraft overtake the slower ones. This is not possible on rails, as a train can overtake another one only at dedicated parts of the tracks. Hence, a single train being delayed can easily paralyze the entire set of trains on the same track. Second, unlike the air traffic environment, the rail system is relatively open. It is difficult, or even impossible, to supervise or restrict unauthorized access to the entire rail infrastructure, while airplanes as well as air traffic control and the communication equipment are located in physically restricted areas.

Equipment on trains and near the tracks is vulnerable to unauthorized access. This poses a security threat. The situation is reminiscent of the challenges of modern mobile and internet networks. However, there is one crucial difference. Tampering with a base station of a mobile network or a Wi-Fi access point may cause service degradation or even unavailability of the service, while tampering with a rail traffic control system can cause serious accidents (e.g., the collision of passenger and chemical trains) or paralysis of rail traffic (i.e., stopping all trains on the track). Such threat examples have the same level of severity as in aviation systems [12]–[15]. There is a good guidebook for cybersecurity planning for railway systems [16] and a description of how cybersecurity requirements and recommendations can be addressed in the railway sector [17].

With all of this in mind, there are two major security challenges to overcome. The first challenge is how to build a tamper-proof traffic control system for a future railway that is deployable in an open operating environment. The second challenge is how to ensure the security measures cannot be utilized as means to generate denial-of-service attacks which could cause the traffic to be paralyzed, as the built-in security measures prevent trains from proceeding on the tracks. Moreover, conventional security measures adopted from other systems (such as air traffic control) are not adequate for an open-access environment, such as railroads.

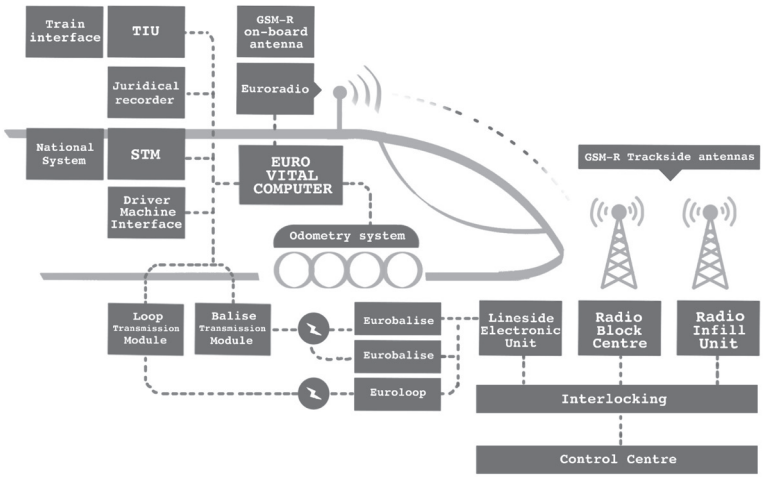
This article focuses on railway communication systems from a European and a future-oriented perspective. It briefly describes the ERTMS standard, its next-generation improvement plans, and related research. Some Asian railway standardization solutions that we consider to be relevant for the European case are also included. Instead of a systematic review of the security area related to railway systems, we focus on two key aspects of cybersecurity, namely cryptography and key management. We anticipate further studies on cybersecurity threats and solutions such as attacks to balise [18] information, wireless on-board equipment, or recorded information and solutions based on artificial intelligence.

2. EUROPEAN RAIL TRAFFIC MANAGEMENT SYSTEM

The European Rail Traffic Management System (ERTMS) [20] is the European standard for automatic train protection (ATP) as well as for command and control systems. It stems from an EU project to unify the European railways and make them more competitive and safer. One of the drivers of this is to facilitate rail traffic between member states. Hence, it constitutes part of the Single European Railway Area [19], which strives to establish nine core network corridors for freight and passengers through European member states. According to the vision, by 2030 these corridors (i.e., 50,000 km of railways) will be equipped with ERTMS capable technology.

The two main building blocks of ERTMS [20]–[22] (Figure 1) are the European Transport Control System Signalling (ETCS) and the Global System for Mobile Communications – Railway (GSM-R). The ETCS consists of both on-board and trackside equipment (Eurobalises [18]). This equipment monitors the environment and the driver and can make a decision to stop the train. The GSM-R utilizes frequency bands separate from those used by the GSM. Specification sets for the ETCS as well as the GSM-R emerge from major versions or baselines of the corresponding standards. Backward compatibility is an important feature in this process.

FIGURE 1: ERTMS SUBSYSTEMS (REDRAWN AFTER [23]); BRIEF SUBSYSTEM DESCRIPTION IN TABLE II



The signaling system, the ETCS, is further divided into five levels [23]–[26] (Table I) and multiple operational modes corresponding to different management situations. The different levels enable on-board as well as trackside equipment with varying functionality levels to interoperate. Thus, the design of the standard has foreseen stepwise upgrades. Physically, an ERTMS system consists of two sets of equipment: (i) one on-board the train and (ii) one on the trackside (see Table II). Not all components are necessarily present in all kinds of systems.

TABLE I: ETCS LEVELS [18], [24]–[26] WHERE STML IS SPECIFIC TRANSMISSION MODULE LEVEL, AND RBC IS RADIO BLOC CENTER

Level 0	ETCS-equipped trains running on tracks which are not equipped with corresponding equipment. Movement authorities by lineside signals.
STML	Trains equipped with ETCS running on tracks equipped with national signalling systems. Use of lineside signals depends on compatibility between systems.
Level 1	Train movement supervision. In order to complete this task, lineside signals are required. It includes information about movement authorities, which enable maintenance of needed breaking distances. Also train integrity—i.e., that all wagons are still connected to the train can be checked. However, this functionality is not part of ERTMS.
Level 2	Continuous communication over GSM-R between train and trackside equipment. At this level, lineside signals are optional. Train integrity checking is beyond the scope of ERTMS. RBC can control the train movement. Eurobalises are used for positioning.
Level 3	Continuous communication between train and trackside. On this level, location and integrity checking is included in the ERTMS scope. Thus, lineside signals are not necessary, as Eurobalises suffice. The train itself supervises train integrity.

TABLE II: ERTMS COMPONENTS [18], [23]

Trackside equipment	Abbreviation	Explanation
Eurobalises		Passive elements residing on the track. They store data related to the infrastructure. Trains passing the balises can read this data.
Lineside Electronic Units	LEU	LEUs interface Eurobalises with interlocking. Here, information flows from interlocking (and lineside signalling) through LEUs to balises, which send these telegrams to the trains.
Euroloops		These provide optional support by filling in information between Eurobalises.
Radio Infill Units	RIU	These provide optional support by improving performance with additional, advance balise information implemented by GSM-R signalling.
Radio Block Centres	RBC	Centralized safety entities which utilize GSM-R connections to have an outlook of the train movement. They can communicate with adjoining RBCs.
Interlockings		Not part of ERTMS. Nevertheless, they have a core role in safe route control as almost all ERTMS structures need an interface between interlocking and ERTMS.
Control Centres		These interconnect all the routes and trains in one area.

On-board components	Abbreviation	Explanation
Euro Vital Computers	EVC	EVCs act as central controllers on board the trains. As such they constitute part of the Automatic Train Protection function.
Driver Machine Interfaces	DMI	DMIs enable drivers to interact with the ETCS. In practice, DMIs are often implemented as a touch screen accepting input from the driver and showing system output to the driver.
Train Interface Units	TIU	TIUs interconnect trains with the ETCSes. Thus, TIUs enable statuses and commands to be sent between trains and ETCSes.
Juridical Recording Units	JRU	JRUs act as storage units recording train data and journeys for later analysis.
Balise Transmission Modules	BTM	BTMs receive and process signals from trackside Eurobalises.
Loop Transmission Modules	LTM	LTMs interconnect trains and Euroloops.
Euroradios		These on-board components cater for the GSM-R communication. Thus, they interconnect trains and tracks as represented by RBCs or RIUs.
Odometry systems		These calculate distance travelled. Depending on the hardware setup they calculate distance, speed and acceleration.
Specific Transmission Modules	STM	STMs act as gateways between the ETCSes and various national systems. Hence, they enable interoperability between national control systems and on-board components of ETCSes. Further, they enable smooth transitions from/to national systems.

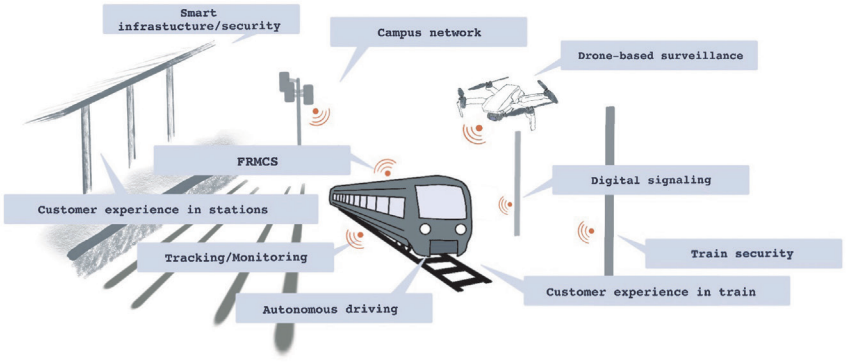
Furthermore, a proposal for a potential Level 4 has emerged outside the ERTMS standard. This additional level relates to static versus dynamic block control. Static block control is how ETCS Levels 1–2 control the physical safety of trains. Static refers to the division of the tracks where trains and train carriages can be located. Static block control may lead to problems if there is heavy congestion on the tracks. On the other hand, dynamic block control refers to a new concept called virtual coupling, which enables dynamic allocation and control of tracks. However, this new concept requires new protocols for communication between railway carriages and sets of carriages. In addition, it is important to remember the consequences of such added functionality. As the virtual train coupling system (VTCS) [27] allows braking distances to become relative, instead of absolute, which saves space on the track, the system becomes more complex and consequently more vulnerable to cyberattacks. Thus, added functionality leads to increased safety and increased security demands. In summary, virtual coupling is not currently part of the ERTMS standard, but extensive studies support its inclusion as an ERTMS Level 4 [28].

3. TOWARDS THE FUTURE

Old, national, non-interoperable railway communication systems need renewal [29]. The current bearer, the GSM-R, provides voice and data communication to trains moving at speeds up to 500 km/h. The GSM-R system builds on 2G GSM cellular technology. Services provided by the GSM-R include group calls with push-to-talk (PTT), voice broadcast, railway emergency calls, prioritization, call pre-emption, and functional and location-dependent addressing. Replacing the GSM-R with GPRS/Edge would enable packet-switched data. However, it would offer only temporary relief as it does not provide enough support for growing ERTMS demands such as autonomous train operation. Furthermore, the availability or support of GSM-R equipment may diminish.

In theory, railway communications could build on many different technology options [29]. Nevertheless, using dedicated bearers means fewer parties sharing networks and costs. This leads to slower deployment. Hence, if railways can build upon shared communication services, the services are cheaper and become available faster. What matters is the total cost of ownership. On the other hand, requirements are growing. Nowadays, train passengers expect a higher level of services, such as advanced guiding services, real-time travel information, and infotainment services. Furthermore, trains move faster, increasing the requirements on the underlying infrastructure. Such expectations heavily load even the current 4G infrastructure. The objective is to provide services for both operators and passengers within the same network. Handovers for both controls and passengers need to be fast and seamless. (Figure 2)

FIGURE 2: EXPECTATIONS OF THE NEXT-GENERATION SYSTEM (REDRAWN AFTER [30]) WHERE FRMCS IS THE FUTURE RAILWAY MOBILE COMMUNICATION SYSTEM



At present, the ERTMS specifies that the GSM-R constitutes its communication carrier. However, shortcomings of the GSM and GSM-R implementations (bandwidth, security, flexibility, and cost) have led to proposals to add newer generations of mobile communications, such as 4G and 5G technologies, to the standard. Within 4G, the main contender is the Long-Term Evolution for Railways (LTE-R) system. On the other hand, the Future Railway Mobile Communication System (FRMCS) offers a 5G based solution.

Asian countries have actively developed LTE-R as a replacement for GSM-R [31]. Likewise, standardization of LTE-R is underway in the European Telecommunications Standards Institute (ETSI) [32]. Although the ultimate goal for the ETCS Level 3 carrier is a 5G-solution, there is still an urgent need to migrate the existing GSM-R-based networks to a more suitable technology. As LTE-R is already production-grade (i.e., commercial off-the-shelf (COTS) technology), it seems like the logical next step towards 5G in Europe too, and particularly for high-speed trains [33]. Train networks based on the Chinese train control system (CTCS)[1] have employed LTE-R since 2016 [34], [35]. Therefore, there is already evidence that such a migration path is also applicable to European networks.

The next-generation development effort [29], the FRMCS, builds on a 5G service standard originally aimed at public safety. However, other areas with mission-critical communication needs have also adopted it, such as railways, shipping, air transport, and industry. This global mission-critical communication standard (MCX) arose under the auspices of the Third Generation Partnership Project (3GPP). Corresponding to the different requirements of the MCX target group, compared to a commercial target group, MCX supports features such as emergency calls, group calls, PTT, stronger security and encryption, as well as higher reliability and availability.

4. SECURITY OF 5G EVOLUTION IN EUROPEAN RAILWAY MANAGEMENT

A. Towards ETCS Level 3

The ERTMS is more of an interface standard than an implementation standard. In addition, a low ETCS application level implies greater dependence on national information systems, which rely on national security solutions. It has been noted [36] that the most serious vulnerabilities may more likely arise from case-dependent security implementations than from known issues in the ERTMS standard. However, as European nations move towards a common standard, it is also ever more important to make that common standard secure.

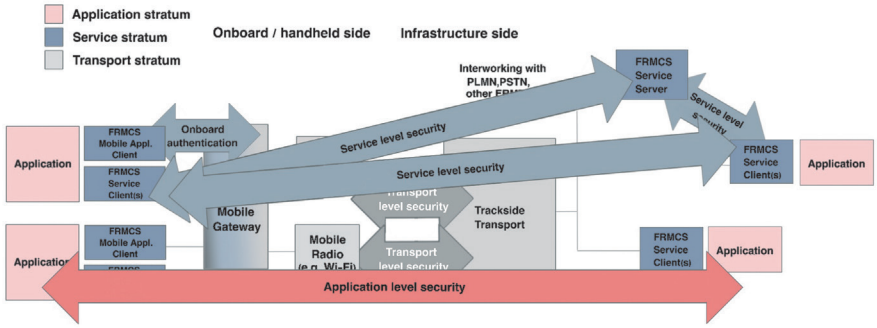
1 CTCS is a very ETCS-like standard for the Chinese railways.

The path from national systems to a European standard varies somewhat by nation: From the Finnish national perspective [37], the ETCS Level 3 standardization is still considered unfinished. Thus, for example, the Finnish rail modernization project Digirail considers Level 2 to be the main goal. Finland managed to replace the requirements for the Level 2 carrier network of ERTMS with the Finnish national terrestrial trunked radio (TETRA) [38] implementation instead of the GSM-R. Thus, multiple different routes to ETCS Level 3 and 5G exist, including LTE-R.

The FRMCS appears to be the most likely 5G standard for ETCS Level 3. Security for the FRMCS is being standardized in parallel in ETSI [11] and in 3GPP [39], although it currently appears that the 3GPP work is more mature. The security features of the FRMCS at this point in time are expressed only as high-level requirements, and key management or encryption is only elaborated at an interoperability level.

According to ETSI, the FRMCS system security is expected to be defined on three layers or levels. There will be one independent security concept per what the standard proposal calls "stratum":[2] one each for the application, service, and transport strata [11], as shown in Figure 3.

FIGURE 3: FRMCS SECURITY ARCHITECTURE (REDRAWN AFTER [11][3])



B. Known ERTMS Shortcomings

The ERTMS standard security features as implemented with the GSM-R and Euroradio [18] protocols are already extensively researched (see, e.g., [40]). In [36] the authors review ETCS Level 2 and point out that the design principles of ERTMS are fail-safe, or "when in doubt, stop the train." This type or principle prevents accidents but

2 Using the ETSI terminology, a "stratum" is a collection of functions and services inherited from the 5G security architecture as follows: 5G "access stratum" corresponds to FRMCS "transport stratum"; "non-access stratum" roughly corresponds to "service stratum."

3 The image has been redrawn from the figure in the referenced original [11, p. 27, Fig 5-5] (i.e., with some texts obscured). The obscured texts are: i) the lowest blue box on the left reads "FRMCS Service Client(s)," ii) the text below the infrastructure side reads "Interworking with PLMN, PSTN, other FRMCS networks, GSM-R," iii) the grey arrow on the right reads "Service level security."

leaves the system as a whole vulnerable to denial-of-service attacks. One of the main vulnerabilities in the security model was considered to be the reliance on the set of GSM standards, and also the implicit trust placed on some entities now considered to be air-gapped, such as internal networks on board trains.

The information security objectives of a rail system are safety- and service-oriented, in that order [36]. Thus, integrity is generally valued higher than service availability, and the value of confidentiality is in general the lowest of these three objectives (depending on the use case and implementation).

The main cryptographic primitive in the Euroradio protocol is ISO-9797 Message Authentication Code (MAC) algorithm no. 3, based on the 3DES[4] standard with the CBC[5] mode and three keys [41]. If the carrier layer is implemented with the GSM-R, it commonly uses the GSM A5/1 stream cipher for encryption of traffic between the train and trackside [40]. Other carrier layers impose other algorithms, such as the TETRA AIE-family ciphers (e.g., the Finnish solution) or LTE-R ciphers AES, SNOW 3G, and ZUC (see footnote 10). The high-level cryptographic keys of Euroradio are commonly distributed manually, while the carrier layer top keys of the hierarchy are pre-placed in the SIM cards, from which additional keys for several purposes are derived or agreed.

The detailed ERTMS security shortcomings reported in the literature are as follows:

(i) ERTMS considers the on-board unit (OBU [18]) to be trusted [36]. This is based on the implicit assumption that the on-board circuitry is air-gapped from the other (not ETCS conformant, public) networks.
(ii) The balise system is trusted [40] as communications between the balise and the OBU contain no cryptographic controls at all. However, the balises can usually be reprogrammed on the spot with mere knowledge of the communications protocol and data formats. This could result in all forms of integrity attacks [42]. The balises are also criticized as being too static to incorporate, for example, regarding temporary speed limits on tracks [37].
(iii) Key management in the ERTMS is primarily based on symmetric keys and relies on high-level pre-shared keys (PSK) [36]. This results in a heavy key management burden (due to large networks and the manual operations involved) [36], [43].
(iv) Managing an X.509-type public key infrastructure (PKI) over the GSM-R is deemed too expensive [43], although defined in Subset 137 of the ERTMS specifications [44].
(v) Key compromise in manually managed key material is considered to be a problem [45], [46], as the key transfer via physical tokens is prone to theft.

4 FIPS PUB 46-3 and RFC 1851
5 Cipher Block Chaining.

(vi) Key authentication in railway environments is also considered to be a general problem [47]: if an infrastructure key (e.g., for the carrier network) is replaced by one from a corrupted node, other nodes might not be able to notice this.
(vii) The combination of GSM-R, Euroradio, and Subset 026 application protocols in ERTMS Level 2 can be exploited to circumvent Euroradio's message authentication with non-negligible probability, bordering on a practical-level attack [48]. All the Subset 026 application-level messages could be forged, including, for example, the movement authority messages [18]. This is currently the only known detailed attack against ERTMS using the original security model of the ETCS designers.

The work in [49] reviews 11 different kinds of metadata privacy-related attacks (location, identity, and service privacy) to general carrier level technologies from 2G to 5G (up to Release 15). The FRMCS does not detail the security model in [39], but all aspects of, for example, service level security, are considered to be part of the security goals, including privacy. Of the attacks listed in [49] we consider those that have consequences other than mere privacy issues in LTE-R and GSM-R, and those that are still unresolved as of 5G Release 16. The motivation for leaving location, identity, and service privacy out of consideration in LTE-R and GSM-R is that the main use cases are for keeping the train safe on the track. Further use cases arise only with the 5G service stratum in the FRMCS. These attacks include the following:

(i) International mobile subscriber identity (IMSI) catching: completely hijacking the carrier-layer connection is naturally disastrous for higher-layer protocol message integrity (discussed below) but also other types of fake base station attacks can be mounted to undermine 5G security in general. If, for example, LTE-R easily reveals additional information about communication identities to an attacker. The full solution to this (a detection architecture) is not purely cryptographic, however.
(ii) Radio resource control (RRC) idle mode security is poorly addressed in 3G and 4G and thus also in LTE-R. This enables rogue base station attacks. (RRC is a radio resource management protocol between the mobile device, such as the train, and the serving network, such as trackside equipment.)
(iii) Raw IMSI probing monitors mobile cell signaling after guessing a mobile station's identity to determine whether that particular station was there or not. This affects location and service privacy but is difficult and expensive to remedy. The attack is slow to implement, which is why it has also been left unaddressed by 3GPP in 5G [49].
(iv) Cell radio network temporary identifiers (C-RNTIs) are physical layer identifiers, whose poor handling within a cell infrastructure may leak the

mobile station's location and service information. The authors of [49] believe that such an attack cannot be mitigated without rolling out a full PKI, which on the physical level would be too expensive.

(v) The authentication protocols within 4G and 5G (e.g., EAP-AKA[6]) have had multiple security and privacy issues (see, e.g., [50]). The main issue considered in [49] is the use of certain status-dependent error codes to leak the states and thus also service information. A full fix for large deployments would again require a full PKI.

The Euroradio protocol has been formally analyzed on several occasions [51]–[54]. A separate formal analysis focusing on security [51] found the following weaknesses related to cryptography:

(i) High-priority messages are not authenticated, enabling, for instance, unauthorized emergency stop messages if the underlying carrier network (e.g., GSM-R) is compromised by, for example, IMSI catching.

(ii) Session establishment in Euroradio communication is possibly vulnerable to downgrading attacks (the standard does not specify how disagreement over security parameter negotiations should be handled in all cases).

(iii) The standard does not mandate for the train to identify RBCs if they all lie within the same security domain (i.e., if they share the high-level keys). Thus, the messages could conceivably be forwarded between multiple RBCs to distort the operational picture of trains on the track.

In addition, Euroradio employs the 3DES block cipher for message authentication codes (MACs), although 3DES has a block size of 64 bits. This implies that forging any message in a chosen-message attack (CMA) security model (free access to MAC verification circuitry) is trivial as the computational power for 2^{64} MAC evaluations is not difficult to achieve. Hence, the block size of the current MAC scheme is a major problem.

C. Cryptographic Improvements in 5G and Implications on ERTMS

All of the cryptography solutions of the mobile communication generations ("Gs") before 5G have been based on symmetric keys, focusing more on network traffic confidentiality (and user identification for billing purposes) than on other aspects of network security (integrity, authentication, metadata privacy). Furthermore, before 5G, the network infrastructure was considered more or less trusted, enabling so-called "IMSI-catcher" attacks, where the connection between the mobile station and base station is hijacked by a man-in-the-middle (MitM).[7] The cryptographic algorithms

6 Extensible Authentication Protocol Authentication and Key Agreement

7 Technically, the later "Gs" no longer allow MitM attacks as severe, but even in LTE some MitM attacks are still possible (see, e.g., [61]). The complete network information with authorizations should also be verified and protected attributes should include, for example, location.

in 2G networks turned out to be insecure, and even in 4G there are some crypto algorithmic solutions as part of the standard that can be considered national more than international, such as ZUC [55], which was developed in China specifically for national use and the Asian market.

The set of 5G standards from 3GPP includes multiple security improvements compared to, for example, 3G and 4G. We list below the key (cryptography-related) improvements we believe are significant for future railways. Other improvements are briefly listed, for example, in [56] and [57] and in the latest full 5G security specification in [58].

The most relevant cryptographic security improvements in 5G are as follows:

(i) Support for asymmetric cryptography, mainly with the Elliptic Curve Integrated Encryption Scheme (ECIES). However, the full scale of a PKI is considered to be outside the scope of 5G; instead, the asymmetric keys are burned to the universal subscriber identity module (USIM) or managed inside operator network nodes [56]. The use of ECIES in concealing the subscription permanent identifier (SUPI) further mitigates privacy-oriented IMSI-catching attacks [49].
(ii) Improved authentication protocols and their user requirements in the form of use cases. In the ERTMS, service authentication at the application level is implemented by the ERTMS Subset 026 protocol, but with 5G, these services can be authenticated by the 5G-infrastructure service in the FRMCS service stratum.
(iii) Cryptographic algorithms intended for integrity protection are based on one-way functions with integrity tags of sufficient length [58].
(iv) An improved key hierarchy (as in more refined levels and specifically purposed keys, enabling support for more security services with proper separation of the keying material) in multiple network node types [58]. Key management still only considers key generation, derivation, and distribution, but no other functionality, such as key-update or over-the-air (OTA) functions [56].
(v) Multiple separate simultaneous security contexts for one item of user equipment and two serving networks enable more secure associations between the OBU and RBCs. This applies to handover scenarios as well. Although, the current 5G standards leave the full security features of the handover implementation optional [56].

(vi) Support for cryptography on multiple protocol layers, from the 5G access stratum to the non-access stratum. The EAP[8] and Internet protocol security (IPsec) based security in the integrated access and backhaul (IAB[9]) nodes (since 5G Release 16) enables a more cost-effective setup of trackside infrastructure and emergency services. Adding integrity protection to all RRC messages removes some types of fake base station attacks [49].

(vii) New industrial internet of things (IIoT) security features (since Release 16) enable secure, fast, and reliable connections, for example, for high-speed trains. Some nations' trackside requirements (such as Finnish, see [59]) call for duplicated connections which are now fully supported by 5G's standard design.

(viii) Multiple new improvements to detect and prevent fake base stations, for example, RRC message integrity protection and a framework to try to detect the fake base stations. The latter improvement mechanism is so far informative only [49] and it is unclear whether it will be incorporated into, for example, the FRMCS.

Algorithm and key downgrading attacks are still possible (e.g., the support for NULL or no encryption and integrity [56]), but newer releases improve this (notably mandatory full-rate user plane integrity protection in Release 16 types [57]).

The security improvements in LTE-R, 5G, and FRMCS discussed above were evaluated against the currently identified shortcomings in ERTMS Level 2 security with the GSM-R. These findings have been collected in Table III.

In Table III, the first column lists the security concerns and the following columns show how a selected carrier technology can mitigate these concerns, starting from the current combination of Euroradio (ER) over GSM-R, moving to ER over LTE-R, and finally to the planned 5G, with the assumed change to FRMCS-like architecture (or at least a fundamental overhaul from Euroradio). Due to the absence of a railway version of 5G (possibly expressed someday in the FRMCS), we based the evaluation on the assumption that the current 5G security potential would be fully embraced by the future railway security standard as well.

8 Extensible Authentication Protocol, IETF RFC 3748 and the respective use in TLS in RFC 5216.

9 IAB is a 5G concept for connecting remote sites to a central facility. The concept is useful for building and connecting trackside equipment in rural areas.

TABLE III: SECURITY CONCERNS

Security concern	ER/GSM-R	ER/LTE-R	FRMCS/5G
Small block length of authentication tags	–	–	+
Lack of modern cryptographic primitives	–	~	+
Lack of cryptographic protection in balises	–	–	~
Heavy key management (resulting from symm. key)	–	–	+
Lack of public-key use and management	–	–	~
Manually managed key material	–	–	+
Key authentication	–	–	+
Network authentication	–	~	+
Service authentication	–	–	+
Downgrading attacks	–	–	+
RBC identification	–	~	+
RBC handover keying	–	~	+
IMSI catching	–	–	+
RRC idle mode security	–	–	+

Legend

– does not mitigate the security concern at all
\+ security concern is fully mitigated
~ whether and tho what extent the security concern is mitigated depends on implementation, case and definition

We elaborate on some of the evaluations depicted in Table III, unless evident from the previous discussion:

(i) The main concern in regard to outdated cryptographic primitives is a problem found in both Euroradio and GSM. In LTE-R implementations, the carrier level is modernized,[10] but judging from the LTE-R adoption in China, it is likely that the Euroradio protocol would not be replaced along with LTE-R in Europe, and the primitives on that level would remain unchanged.

(ii) Protection of balise communication is not dependent on the carrier technology. The table reflects the situation at the higher protocol level

10 This is based on the LTE standard cryptographic algorithms EEA1-3 and EIA1-3 (AES, SNOW3G and ZUC, respectively). However, as LTE-R is not standardized internationally or in Europe, it is difficult to tell which of these algorithms will make their way into implementations. One example is the BEEHD LTE-R product by Softil, which claims to use AES [62], [63].

(ETCS Level 2), while 5G is assumed to be connected to ETCS Level 3. However, with the shift to 5G technologies and the FRMCS, the location service offered by balises is proposed to be augmented or even replaced by a combination of satellite location services (like GPS) and 5G location information, such as cell identification [26], [37].

(iii) Key management with ER/LTE-R is not any simpler than with ER/GSM-R. The required manual key management for the Euroradio layer would not change, and LTE-R key management was criticized as being heavy [35].

(iv) Key authentication in general and especially in RBC-OBU connection handovers was considered a problem. This has not yet been remedied in LTE-R [35]. However, the 5G network-user connection is no longer based solely on PSKs, but on an authenticated key agreement scheme based on the ECIES [58]. The RBC-OBU connection security is not yet specified for 5G at this level, but the key management architecture of 5G offers a secure way to do this.

(v) The evaluation of network authentication refers to how well the carrier technology defends itself against IMSI catching. This is markedly difficult with 4G and even more so in 5G, compared to 2G.

(vi) Application-level service authentication in ETCS Level 2 is based on the (vulnerable) Euroradio protocol, but in 5G this can be moved to the 5G non-access stratum authentication mechanisms.

(vii) Downgrading attacks in the current ERTMS is an issue for the Euroradio protocol. Thus, ER/LTE-R cannot mitigate this part. In 5G, the conformance requirement for all protocol levels states that bidding-down attacks must be prevented [58, Clause 5.1.1].

(viii) RBC identification is a problem on the Euroradio layer and on the tight coupling of the application layer to the carrier layer. According to both 5G and the FRMCS, these layers have now been decoupled. LTE-R is not free of the Euroradio protocol, but the use of EPS-AKA[11] on the NAS[12] and RRC[13] levels mitigates the problem somewhat.

(ix) Handover keying is no longer much of a problem in LTE, but it still lacks, for example, forward secrecy [35]. In 5G, the most common handover protocols have been formally analyzed and proven secure within their security model [60].

In summary, an ERTMS evolution, which only replaces GSM-R with LTE-R, does not bring significant security benefits, although the most prominent issues can be solved (i.e., the known MAC-vulnerability and low-level connection hijacking with full IMSI catching). However, a full overhaul of ERTMS to an FRMCS-like architecture would at least more or less remedy the known ERTMS Level 2 security issues. It should also

11 EPS-AKA = Evolved Packet System Authenticated Key
12 NAS = Non-Access Stratum
13 RRC = Radio Resource Control (protocol).

be noted that the 5G architecture employs a large number of additional services (e.g., virtual coupling, VTCS) which also need to be secured. Many of these are already being considered, albeit implicitly, with the 5G Release 16 IIoT security measures.

5. CONCLUSIONS

This paper provided a glimpse of plans and challenges emerging during the standardization of the communication system for the European railways. This article included a brief presentation of the ERTMS standard and the FRMCS. The study highlighted challenges and refinements in two key aspects of cybersecurity, namely cryptography and key management, in proposed railway communication systems.

REFERENCES

[1] International Union of Railways (UIC), "Passenger-tonne-line-kilometers timeseries over period 2004–2019." Accessed: Jan. 4, 2022. [Online]. Available: https://uic.org/IMG/pdf/passenger-tonne-line-kilometers-timeseries-over-period-2004-2019.pdf

[2] World Bank. "Air transport, passengers carried | Data." The World Bank. https://data.worldbank.org/indicator/IS.AIR.PSGR (accessed Jan. 4, 2022).

[3] World Bank. "Air transport, freight (million ton-km) | Data." The World Bank. https://data.worldbank.org/indicator/IS.AIR.GOOD.MT.K1 (accessed Jan. 4, 2022).

[4] World Bank. "Railways, passengers carried (million passenger-km) | Data." The World Bank. https://data.worldbank.org/indicator/IS.RRS.PASG.KM (accessed Jan. 4, 2022).

[5] World Bank. "Railways, goods transported (million ton-km) | Data." The World Bank. https://data.worldbank.org/indicator/IS.RRS.GOOD.MT.K6 (accessed Jan. 4, 2022).

[6] OECD. "Passenger transport." OECD Data. https://data.oecd.org/transport/passenger-transport.htm#indicator-chart (accessed Jan. 4, 2022).

[7] OECD. "Freight transport." OECD Data. https://data.oecd.org/transport/freight-transport.htm#indicator-chart (accessed Jan. 4, 2022).

[8] European Environment Agency. "Transport and environment report 2020: Train or plane?" European Environment Agency. https://www.eea.europa.eu/publications/transport-and-environment-report-2020 (accessed Dec. 30, 2021).

[9] World Bank. "Rail lines (total route-km) – United Kingdom, Spain, Poland, Italy, France." The World Bank | Data. https://data.worldbank.org/indicator/IS.RRS.TOTL.KM?locations=GB-ES-PL-IT-FR (accessed Jan. 4, 2022).

[10] FRMCS Functional Working Group, "Future Railway Mobile Communication System User Requirements Specification," Feb. 2020.

[11] ETSI, "TR 103 459 - V1.2.1 - Rail Telecommunications (RT); Future Rail Mobile Communication System (FRMCS); Study on system architecture," 2020.

[12] G. Lykou, G. Iakovakis, and D. Gritzalis, "Aviation Cybersecurity and Cyber-Resilience: Assessing Risk in Air Traffic Management," in *Critical Infrastructure Security and Resilience*. Cham, Switzerland: Springer, 2019, pp. 245–260.

[13] A. W. Evans, "Fatal train accidents on Europe's railways: An update to 2019," *Accid. Anal. Prev.*, vol. 158, Aug. 2021, doi: 10.1016/j.aap.2021.106182.

[14] M. Khanmohamadi, M. Bagheri, N. Khademi, and S. F. Ghannadpour, "A security vulnerability analysis model for dangerous goods transportation by rail – Case study: Chlorine transportation in Texas-Illinois," *Saf. Sci.*, vol. 110, pp. 230–241, Dec. 2018, doi: 10.1016/j.ssci.2018.04.026.

[15] J. L. Wybo, "Track circuit reliability assessment for preventing railway accidents," *Saf. Sci.*, vol. 110, pp. 268–275, Dec. 2018, doi: 10.1016/J.SSCI.2018.03.022.

[16] International Union of Railways (UIC), *Guidelines for Cyber-Security in Railway*. UIC-ETF, 2018.

[17] *Railway applications - Cybersecurity*, CLC/TS 50701:2021, European Committee for Electrotechnical Standardization, 2021, pp. 1–161.
[18] ERA * UNISIG * EEIG ERTMS USERS GROUP, "Glossary of Terms and Abbreviations," 2016. Accessed: Mar. 9, 2022. [Online]. Available: https://www.era.europa.eu/sites/default/files/filesystem/ertms/ccs_tsi_annex_a_-_mandatory_specifications/set_of_specifications_3_etcs_b3_r2_gsm-r_b1/index003_-_subset-023_v330.pdf
[19] European Council. "Building the single European railway area." European Council. https://www.consilium.europa.eu/en/policies/single-eu-railway-area/ (accessed Jan. 7, 2022).
[20] European Commission. "ERTMS: What is ERTMS about?" European Commission. https://transport.ec.europa.eu/transport-modes/rail/ertms_en (accessed Oct. 15, 2021).
[21] European Commission. "How does it work?" European Commission. https://ec.europa.eu/transport/modes/rail/ertms/how-does-it-work_en (accessed Oct. 15, 2021).
[22] European Commission. "Set of Specifications and Baselines." European Commission. https://transport.ec.europa.eu/transport-modes/rail/ertms/how-does-it-work/set-specifications-and-baselines_en (accessed Oct. 15, 2021).
[23] European Commission. "Subsystems and Constituents of the ERTMS." European Commission. https://transport.ec.europa.eu/transport-modes/rail/ertms/how-does-it-work/subsystems-and-constituents-ertms_en (accessed Oct. 15, 2021).
[24] European Commission. "ETCS Levels and Modes." European Commission. https://transport.ec.europa.eu/transport-modes/rail/ertms/how-does-it-work/etcs-levels-and-modes_en (accessed Oct. 15, 2021).
[25] A. El Amraoui and K. Mesghouni, "Performing enhanced rail formal engineering constraints traceability: Transition modes," in *2015 International Conference on Industrial Engineering and Systems Management (IESM)*, Oct. 2015, pp. 61–66, doi: 10.1109/IESM.2015.7380136.
[26] European Union Agency for Railways. "Set of specifications 3 (ETCS B3 R2 GSM-R B1) | ERA." https://www.era.europa.eu/content/set-specifications-3-etcs-b3-r2-gsm-r-b1_en (accessed Jan. 7, 2022).
[27] Shift2Rail. "X2RAIL-3." Shift2Rail. https://projects.shift2rail.org/s2r_ip2_n.aspx?p=X2RAIL-3 (accessed Jan. 7, 2022).
[28] E. Goddard *et al.*, "ERTMS Level 4, Train Convoys or Virtual Coupling," 2016. [Online]. Available: https://webinfo.uk/webdocssl/irse-kbase/ref-viewer.aspx?Refno=1882928268&document=ITC Report 39 Train convoys and virtual coupling.pdf
[29] Carmen Patrascu *et al.*, "5G VICTORI Use case and requirements definition and reference architecture for vertical services; Draft 1.0," 2020. Accessed: Dec. 30, 2021. [Online]. Available: https://www.5g-victori-project.eu/project-outcomes/deliverables/
[30] 3G4G Blog. "Future Railway Mobile Communication System (FRMCS)." The 3G4G Blog. https://blog.3g4g.co.uk/2021/09/future-railway-mobile-communication.html (accessed Jan. 8, 2022).
[31] Andrzej Kochan and Łukasz Gruba, "Analysis of the Migration Process in the ERTMS System from GSM Technology to LTE on the Polish Railway," in *Management Perspective for Transport Telematics*, J. Mikulski, Ed. Cham, Switzerland: Springer International Publishing, 2018, pp. 249–262.
[32] *TS 122 289 - V16.1.0 - LTE; 5G; Mobile communication system for railways (3GPP TS 22.289 version 16.1.0 Release 16)*, TSGS, 2020.Accessed: Dec. 1, 2021. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx
[33] R. He *et al.*, "High-Speed Railway Communications: From GSM-R to LTE-R," *IEEE Veh. Technol. Mag.*, vol. 11, no. 3, pp. 49–58, Sep. 2016, doi: 10.1109/MVT.2016.2564446.
[34] Railway Technology. "Huawei introduces LTE-R Solution for wireless rail communications." Railway Technology. https://www.railway-technology.com/news/huawei-introduces-lte-r-solution-for-wireless-rail-communications/ (accessed Jan. 7, 2022).
[35] Y. Wang, W. Zhang, X. Wang, W. Guo, M. K. Khan, and P. Fan, "Improving the Security of LTE-R for High-Speed Railway: From the Access Authentication View," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–15, 2020, doi: 10.1109/TITS.2020.3024684.
[36] R. R. Bloomfield, R. Bloomfield, I. Gashi, and R. Stroud, "How secure is ERTMS?" in *International Conference on Computer Safety, Reliability, and Security*, 2012, pp. 247–258.
[37] Liikenne- ja viestintäministeriö, "Kohti digitaalista ja älykästä rautatieliikennettä: Digirata-selvityksen loppuraportti," 2020.
[38] ETSI. "TETRA | TErrestrial Trunked RAdio." ETSI. https://www.etsi.org/technologies/tetra (accessed Mar. 9, 2022).
[39] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Study on Future Railway Mobile Communication System; Stage 1 (Release 17)," 2021. Accessed: Jan. 7, 2022. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3162

[40] R. J. Thomas, "A systematic development of a secure architecture for the European Rail Traffic Management System," Ph.D. dissertation, School of Comput. Sci., Univ. of Birmingham, Birmingham, UK, 2019.
[41] *Information technology — Security techniques — Message Authentication Codes (MACs) — Part 1: Mechanisms using a block cipher*, ISO/IEC 9797-1:2011, ISO, 2011, p. 11.
[42] H. Wei Lim *et al*., "Data Integrity Threats and Countermeasures in Railway Spot Transmission Systems," *ACM Trans. Cyber-Physical Syst.*, vol. 4, no. 7, 2019, doi: 10.1145/3300179.
[43] R. J. Thomas, M. Ordean, T. Chothia, and J. de Ruiter, "TRAKS: A universal key management scheme for ERTMS," in *Proceedings of the 33rd Annual Computer Security Applications Conference*, Dec. 2017, pp. 327–338, doi: 10.1145/3134600.3134631.
[44] UNISIG, "On-line Key Management FFFIS Company Technical Approval Management approval," Dec. 17, 2015. Accessed: Jan. 7, 2022. [Online]. Available: https://www.era.europa.eu/sites/default/files/filesystem/ertms/ccs_tsi_annex_a_-_mandatory_specifications/set_of_specifications_3_etcs_b3_r2_gsm-r_b1/index083_-_subset-137_v100.pdf
[45] X. Hei, W. Gao, Y. Wang, Z. Liang, W. Ji, and X. Hu, "Railway Key Exchange Scheme for Improving Communication Efficiency of RSSP-II Protocol," in *2019 IEEE Globecom Workshops* (GC Wkshps), Dec. 2019, pp. 1–6, doi: 10.1109/GCWkshps45667.2019.9024450.
[46] L. Zhang, J. Bai, and P. Jiang, "Research on Key Management Scheme of X2 Handover Protocol in LTE-R," in *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, Dec. 2019, pp. 1479–1483, doi: 10.1109/ICCC47050.2019.9064161.
[47] S.-Y. Chang, S. Cai, H. Seo, and Y.-C. Hu, "Key Update at Train Stations: Two-Layer Dynamic Key Update Scheme for Secure Train Communications," in *Security and Privacy in Communication Networks. SecureComm 2016. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, R. Deng, J. Weng, K. Ren, V. Yegneswaran, Eds. 2017, vol 198, pp. 125–143. doi: 10.1007/978-3-319-59608-2_7.
[48] T. Chothia, M. Ordean, J. de Ruiter, and R. J. Thomas, "An Attack Against Message Authentication in the ERTMS Train to Trackside Communication Protocols," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, Apr. 2017, pp. 743–756, doi: 10.1145/3052973.3053027.
[49] H. Khan and K. M. Martin, "A survey of subscription privacy on the 5G radio interface – The past, present and future," *J. Inf. Secur. Appl*., vol. 53, Aug. 2020, Art. no. 102537, doi: 10.1016/j.jisa.2020.102537.
[50] J. Munilla, M. Burmester, and R. Barco, "An enhanced symmetric-key based 5G-AKA protocol," *Computer Networks*, vol. 198, Oct. 2021, Art. no. 108373, doi: 10.1016/j.comnet.2021.108373.
[51] J. de Ruiter, R. J. Thomas, and T. Chothia, "A Formal Security Analysis of ERTMS Train to Trackside Protocols," in *Reliability, Safety, and Security of Railway Systems. Modelling, Analysis, Verification, and Certification*, Th. Lecomte, R. Pinger, and A. Romanovsky, Eds. Cham, Switzerland: Springer International Publishing, 2016, pp. 53–68.
[52] E. Rosaria, L. Armando, M. Pietro, and S. Angela, "Formal verification of ertms euroradio safety critical protocol," *Proceedings 4th Symposium on Formal Methods for Railway Operation and Control Systems (FORMS'03)*, 2003.
[53] L. Hongjie, C. Lijie, and N. Bin, "Petri Net-based Analysis of the Safety Communication Protocol," *TELKOMNIKA Indones. J. Electr. Eng*., vol. 11, no. 10, Oct. 2013, doi: 10.11591/telkomnika.v11i10.3462.
[54] Y. Zhang et al., "Formal verification of safety protocol in train control system," *Sci. China Technol. Sci*., vol. 54, no. 11, pp. 3078–3090, Nov. 2011, doi: 10.1007/s11431-011-4562-2.
[55] ETSI/SAGE, "Specification of the 3GPP Confidentiality and Integrity Algorithms 128-EEA3 & 128-EIA3. Document 2: ZUC Specification," Jun. 2011. Accessed: Jan. 7, 2022. [Online]. Available: https://www.gsma.com/aboutus/wp-content/uploads/2014/12/eea3eia3zucv16.pdf
[56] R. Piqueras Jover and V. Marojevic, "Security and Protocol Exploit Analysis of the 5G Specifications," *IEEE Access*, vol. 7, pp. 24956–24963, 2019, doi: 10.1109/ACCESS.2019.2899254.
[57] M. Wifvesson and P. K. Nakarmi. "A summary of 3GPP release 16, 5G phase 2: Security and RAN." Ericsson. https://www.ericsson.com/en/blog/2021/4/3gpp-release-16-5g-phase-2-security-ran (accessed Jan. 7, 2022).
[58] *TS 133 501 - V16.3.0 - 5G; Security architecture and procedures for 5G System (3GPP TS 33.501 version 16.3.0 Release 16)*, TSGS, 2020. Accessed: Jan. 7, 2022. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx
[59] Finnish Transport Infrastructure Agency, "5G in the activities of the Finnish transport infra-structure agency FTIA as a user and enabler of fast data connections," *FTIA Publ*., 2019.
[60] A. Peltonen, R. Sasse, and D. Basin, "A comprehensive formal analysis of 5G handover," in *Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, Jun. 2021, pp. 1–12, doi: 10.1145/3448300.3467823.

[61] A. Shaik, R. Borgaonkar, N. Asokan, V. Niemi, and J.-P. Seifert, “Practical Attacks Against Privacy and Availability in 4G/LTE Mobile Communication Systems,” 2016, doi: 10.14722/ndss.2016.23236.
[62] Softil. “Daeyoun Selects Softil Technology for New Line of MCC Terminals.” Softil. https://www.softil.com/press-release/softil-daeyoun-pr-091818/ (accessed Jan. 8, 2022).
[63] Softil. “BEEHD Client Framework for Mission Critical Communications over LTE/5G.” Softil. https://www.softil.com/wp-content/uploads/2021/05/BR-BEEHD-client-framework-for-Public-Safety-over-LTE_RevC_web.pdf (accessed Jan. 8, 2022).

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

Security and Privacy Issues of Satellite Communication in the Aviation Domain

Georg Baselt
ETH Zurich
Department of Computer Science
Zurich, Switzerland
gbaselt@student.ethz.ch

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

James Pavur
University of Oxford
Department of Computer Science
Oxford, United Kingdom
james.pavur@cs.ox.ac.uk

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

Ivan Martinovic
University of Oxford
Department of Computer Science
Oxford, United Kingdom
ivan.martinovic@cs.ox.ac.uk

Abstract: Modern aviation systems increasingly use satellite channels for data communication. However, many SATCOM providers do not offer encryption below the application layer by default, making their services vulnerable to eavesdroppers and creating security concerns. This research analyses such vulnerabilities specifically with regard to the aviation domain.

We show that even low-resourced attackers can exploit this lack of security. We capture a broad range of SATCOM transmissions in the Ku-Band frequencies using a TV Tuner Card and widely available low-budget equipment for under 400 US dollars. Over 370 GB of aviation-related satellite-downstream data from high-throughput satellites were analysed from a measurement site in Central Europe.

The results of this campaign reveal both security and privacy concerns across the whole spectrum of the industry. We identify unencrypted SATCOM usage comprising usage from in-flight entertainment systems to leaked private encrypted keys. Furthermore, we identified 328 specific aircraft broadcasting their live operations, including three government aircraft that actively blocked any information on their flights from air-traffic tracking sites.

This work concludes with recommendations for both satellite service providers and aviation stakeholders on how these issues could be solved by using encryption at different network layers.

Keywords: *aviation, satellite security, privacy, communication*

1. INTRODUCTION

The aviation industry is one of the world's largest and most important transportation businesses, carrying billions of passengers every year. The International Civil Aviation Organization ICAO estimated that in 2019 alone, a total of 4.5 billion passengers travelled by aircraft, an increase of almost 1.7 billion compared to just ten years earlier [1]. Their projections from 2019 foresaw a total increase to 10 billion passengers per year worldwide by the year 2040. Although this estimated rapid growth has slowed due to the COVID-19 pandemic, the trend remains clear.

A key technology enabling this growth is the usage of satellite communications (SATCOM). Satellite channels allow for fast message transfers in hard-to-reach areas such as oceans, where other communication methods cannot be used. Therefore, they offer reliable bandwidth and higher speeds for both entertainment and safety-critical systems compared to traditional air-to-ground links.

While SATCOM enables aircraft to be as connected as never before, it also introduces new risks and challenges. In particular, new concerns about privacy and data security have emerged in recent years [2]. Advancements in consumer technology saw the introduction of software-defined radios, which enabled its users to intercept aviation transmissions. This practically removed barriers to entry, as the necessary equipment to eavesdrop on aviation and satellite channels used to require specialist equipment [3].

This paper illustrates the prevalence of safety and privacy issues within the current satellite communication landscape in the aviation domain. We conduct the first study of aviation-related satellite transmissions using widely available low-cost equipment.

The contributions of this work are as follows:

i. We identify and map geostationary satellites that are used for aviation data link transmissions from a real-world dataset.
ii. We analyse aviation-related SATCOM transmissions and their impact on safety and privacy.
iii. We discuss the results and the implications of such vulnerabilities for the aviation domain and propose potential countermeasures.

The paper is structured as follows. First, background information on communication methods in the aviation industry as well as previous research efforts is given in Section 2. Section 3 explains the experimental setup and methods used to gather data, while Section 4 describes all relevant findings from the experiment. The results are then discussed in Section 5, before Section 6 concludes.

2. BACKGROUND

Communication plays a vital role in managing modern air traffic worldwide. A wide range of messages are sent from and to aircraft to ensure safe and efficient travel in the skies. While early communication systems provided simple voice communication channels from air to ground and vice versa, nowadays messaging services can send and receive automated information about optimal flight routes, positioning information, real-time weather reports and more. On top of the more complex air traffic control (ATC) messages, airlines have identified in-flight internet connectivity as a strong interest of a changing generation of customers. A report by Panasonic states that 'millennials become the largest air travel spending segment by 2025' [4]. According to the same report, '… one in three passengers are choosing airlines based on connectivity and quality of network service.' This development leads to aircraft sending and receiving more messages than ever before. The following chapter aims to provide a brief overview of current communication methods and their usage in the aviation domain.

A. Communication Methods in the Aviation Domain

There are three main categories of communication methods used in the aviation industry today. Figure 1 shows a simplified overview of these systems and their usage in a few selected applications – the Aircraft Communications, Addressing and

Reporting System (ACARS) and the Controller Pilot Data Link Communications (CPDLC) service.

1) Voice Communication

Today's standard method of air-to-ground communication is by voice broadcast over radio frequencies from 3 MHz to 300 MHz, known as VHF and HF. The development of this technology dates as far back as the 1920s and is still the backbone of modern air traffic control (ATC). VHF communication is used to manage densely populated airspaces, where their line-of-sight limited range plays no significant role. Voice communication over HF offers nearly worldwide coverage, even in polar or oceanic regions, but comes at the cost of the signal-to-noise ratio being dependent on atmospheric conditions, which makes it an unreliable choice for handling time-critical ATC messages [6]. As voice communication channels become increasingly congested in areas with high air traffic intensity, there are multiple avenues to shift ATC from voice to datalink channels.

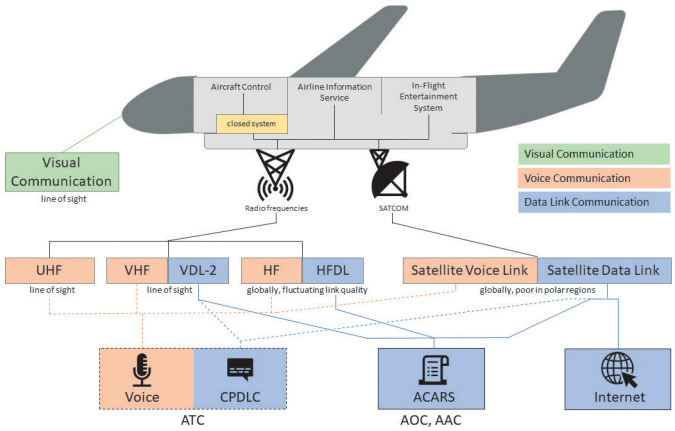
2) Datalink Communication

Datalink communication consists of exchanging digital messages between air and ground. The first system using data links was the Aircraft Communications, Addressing and Reporting System (ACARS) from 1978. It was initially developed to reduce the workload of radio control personnel and to automatically send messages about OOOI[1] events, which informed the receiver about the exact timestamps when the aircraft entered a new major flight phase. Nowadays, ACARS is used to transmit a wide range of clear-text messages to different aviation industry stakeholders, such as Aeronautical Operational Control (AOC) or Airline Administrative Control (AAC) messages to the ground base control of airlines [6].

Another datalink communication system is the Controller Pilot Data Link Communication (CPDLC) service. It serves as a supplementary ATC messaging channel to voice communication and allows its users to send preformatted ATC messages for non-time-critical requests, which significantly reduces the risk of communication errors. Like voice communication, datalink messages can be sent using radio frequencies or alternatively over a satellite connection.

1 Out of the gate, off the ground, on the ground, into the gate.

FIGURE 1: ABSTRACT VIEW OF COMMUNICATION METHODS USED WITHIN THE AVIATION INDUSTRY

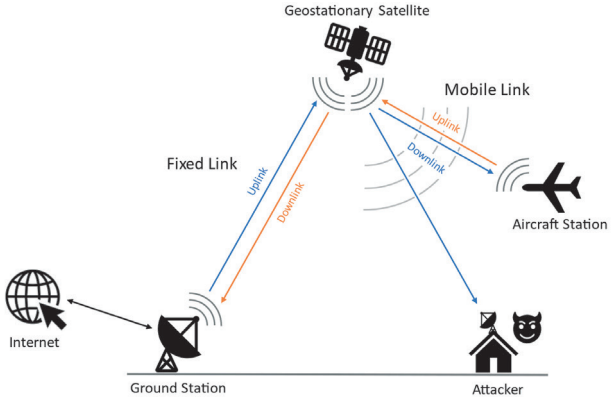


3) Satellite Communication (SATCOM)

For aircraft to use bidirectional SATCOM links, they must be equipped with an SDU (satellite data unit), an antenna and a high-power amplifier. These devices enable aircraft to send messages over radio frequencies via an uplink connection to a satellite, which then relays the received message stream from the aircraft to a ground station. SATCOM datalinks are a two-way communication system, as messages can also be sent from the ground station to the aircraft using the satellite as an intermediate step that broadcasts messages back to the aircraft. The system is depicted in Figure 2 for the case of geostationary (GEO) satellites, which we examine in the present work.

GEO downlink broadcasts can freely be recorded by everybody in the satellite coverage area with the right technical equipment. Messages sent in any other direction could potentially also be listened to, but this would require the eavesdropper to be near the satellite or ground station, as these beams tend to be directed towards their target and are narrower.

FIGURE 2: EAVESDROPPING MODEL ON SATELLITE DOWNLINK COMMUNICATION



Historically, the necessary equipment to receive satellite downlink broadcasts was expensive and difficult to acquire. This effectively acted as a protective barrier for ATC technologies, which have often not been developed with security in mind. ACARS, especially Plain-Old-ACARS, for example, is now used for far more than originally intended and has no default encryption scheme. This barrier-of-entry vanished when software-defined radios (SDR) and applications built on them became widely available in recent years. Attackers are now able to eavesdrop on unencrypted, aviation-related SATCOM feeds using relatively cheap and publicly available equipment.

B. Previous Research

More than a decade ago, Sampigethaya *et al*. first aimed to raise awareness regarding the transition towards fully interconnected flights, such as handling air traffic management over IP [5]. Recently, several works have examined concrete security and privacy issues both in non-satellite aviation datalinks and in satellite networks in different transport domains.

Smith *et al*. [6] use recordings of traditional radio frequencies and SATCOM feeds to illustrate the strong privacy concerns for passengers and crew. With regard to cyber security problems in aviation, research has greatly expanded over the past ten years, covering the full range of communication technologies used by different types of aircraft. For a full survey of these aspects, the reader is referred to [7]. The possible impact of cyber security attacks on safety in aviation has been studied in simulators, indicating the potential for severe disruption [8]. A survey by Strohmeier *et al*. [9], examines the missing awareness from stakeholders inside the aviation industry about such cyber security issues.

Bernsmed *et al*. [10] conducted a risk analysis on the security of future aviation-related SATCOM datalink services. They found severe security concerns in the studied services, where, in some cases, security issues were in direct conflict with safety requirements. The practicability to exploit such security and safety risks was shown by Santamarta in [11]. The authors illustrate the ability of an attacker 'to disrupt, intercept or modify non-safety communications such as InFlight Wi-Fi' as well as 'to attack crew and passengers' devices'. That the use of unencrypted air traffic control links is as insecure in space as it is on the ground has been discussed by the authors in [12]. Finally, a study similar to ours but for the maritime domain was conducted in [3]. The authors illustrate how the unencrypted nature of satellite communication impacts the security of ships around the globe.

To the best of our knowledge, there has been no study where possible security and privacy issues within general-purpose SATCOM data streams used by aircraft have been analysed.

3. METHODS

This section gives an overview of our approach and covers the methods used as well as the technical equipment, both hardware and software.

A. Experimental Design

The goal of our study is to capture aviation-related communication to analyse it for potential safety or privacy issues. Downlink SATCOM messages from geostationary satellites can be received using low-level equipment and a wide footprint area.

Finding such satellites was the first step. We use an exploratory approach, where all geostationary satellites providing coverage at the reception site were identified using an online database. With their coordinates known, the satellite dish was then aligned to receive transmissions from these satellites.

These beams were then scanned for transmissions within specific frequencies and encoding methods. Once a data stream featuring suitable encoding was found, a sample was recorded as a video transport stream file. These files were then analysed by searching for any ASCII encoded aviation-related strings in their byte-code representation.

B. Hardware Setup

Our study followed the setup used in [3] for analysing SATCOM in the maritime domain. The required hardware consists of a satellite dish, a satellite TV tuner card

and a computer connected to it. By using a TV Tuner Card over a software-defined radio, we are better able to keep up with real-time data due to the faster demodulation capabilities of the TV Tuner. It is assumed that an eavesdropper already has access to the latter, which brings the total equipment cost down to under 400 US dollars, as can be seen in Table I.

TABLE I: HARDWARE EQUIPMENT AND COSTS [3]

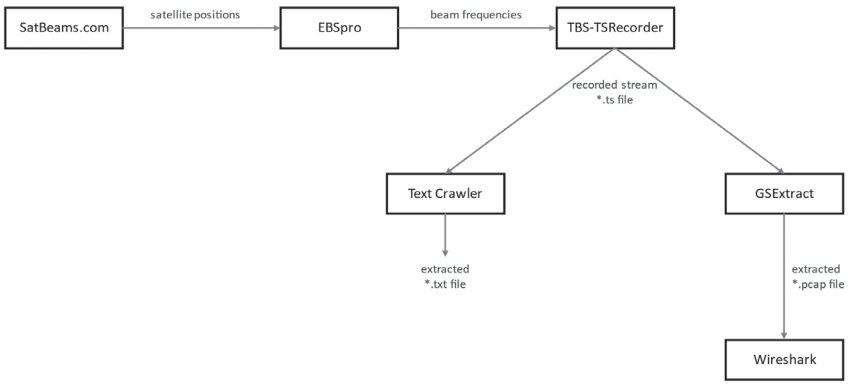
Item	Approximate Cost
TBS-6903 DVB-S2X PCI Card	$300
Selfsat H30D Satellite Dish	$88
3-meter Coaxial Cable	$5
Total	$393

The satellite dish model used for this project is widely commercially available and was combined with a professional-level digital satellite TV tuner card that supports all current digital video broadcasting standards over satellite (DVB-S). Located in Central Europe, with a size of only 517 × 277 × 58 mm, it can receive satellite feeds from all around Europe and parts of the Atlantic Ocean.

C. Software and Methodology

On top of the described hardware, we used open software tools and information in addition to our custom-developed software. Figure 3 provides an overview of the toolchain used for the study.

FIGURE 3: OUTLINE OF THE METHODOLOGY AND SOFTWARE USED TO CAPTURE SATCOM FEEDS



1) Obtaining Satellite Positions

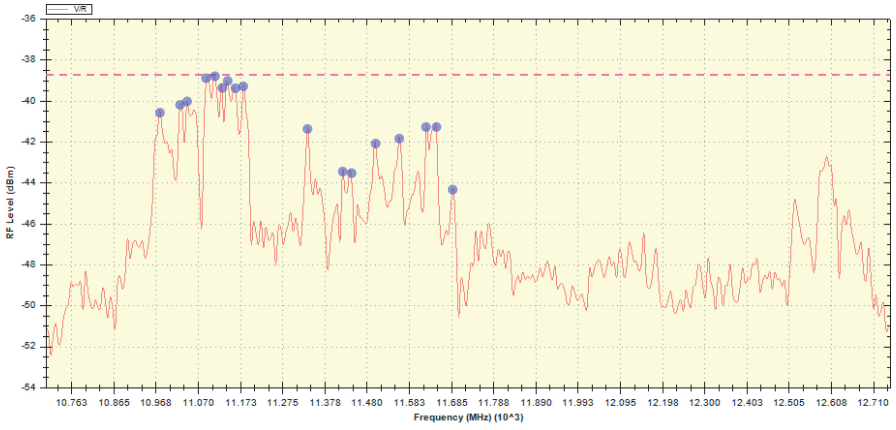
There are publicly available websites that host well-documented databases of satellites and their respective coverage areas. This project used satellite footprint data from the website SatBeams.[2] All geostationary satellites in positions from 30° West to 30° East, covering Central Europe and with beams in the Ku-Band (i.e. in the radio spectrum from 10700 MHz to 12750 MHz) were identified and used for further examination. The sixty-degree window was chosen because of the limitations of the satellite dish, as the signal of satellites farther away would be too weak to capture with the size of the dish used.

The Ku-Band frequencies were chosen, as previous research showed the presence of satellite downstream transmissions in this frequency range. [3]

2) Scanning for Frequencies

These satellite beams were scanned for data streams with a radio frequency using EBS-Pro.[3] We then identified possible data streams through spikes in the signal strength within these frequencies (Figure 4). This scan revealed the frequencies, symbol rates and other encoding meta-data of all satellite streams in this frequency range.

FIGURE 4: THE SIGNAL STRENGTH OF A SATELLITE BEAM IN THE KU-BAND FREQUENCIES. THE MARKED SPIKES REPRESENT POTENTIAL DATA STREAMS AND THE DASHED LINE BENCHMARKS, THE STRONGEST OBSERVED SIGNAL



3) Recording of Streams

Only a subset of all streams featuring some specific encoding schemes and protocols were used for this study. While it would be possible to scan for beams outside the Ku-Band range or with other encoding parameters with the described setup, this could

2 https://satbeams.com
3 https://ebspro.net

require different tools to analyse the recorded data, and therefore was not attempted. Nevertheless, it is reasonable to assume that the selected subset of streams is still representative of a wide range of SATCOM channels, as the chosen parameters are commonly used in practice.

The restrictions for the stream encoding parameters were:

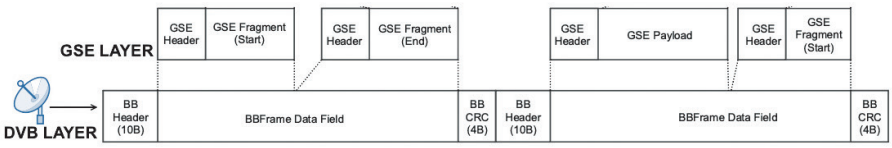
i. Only DVB-S2 streams with Adaptive Coding and Modulation (ACM) and Generic Stream Encapsulation (GSE) were considered.
ii. Only continuous data streams containing packets with MATYPE[4] headers of 42 00 or 43 00 were considered.
iii. Only streams using vertical polarization were considered.

From all the previously identified streams that met these conditions, an initial recording of 500 MB was saved. Later, additional, larger recordings of promising streams were conducted to obtain a larger sample.

4) Data Analysis

As the equipment comprised accessible, low-cost, consumer-grade hard- and software, the recordings can easily be fragmented or corrupted. To extract meaningful data from the lossy feeds, two different methods can be employed: one works on the DVB layer directly and the other targets the higher GSE layer (see Figure 5 for details on the protocol stack).

FIGURE 5: THE LOWER PROTOCOL STACK OF DVB-S COMMUNICATIONS, COMPRISING THE DVB AND THE GSE LAYER



a) GSExtract

For certain DVB-S2 Formats, the forensic tool GSExtract[5] is able to fully recover at least 40% [3] of the recorded GSE packets and convert them into more easily accessible IP traffic files (PCAP). This format can then be processed by Wireshark,[6] a tool for network protocol analysis. GSExtract loops through the recorded GSE-encapsulated files, searching for non-corrupted headers, and uses the information stored in the header to piece IP packets back together.

4 ETSI EN 302 307-1 V1.4.1 (2014-07)
5 https://github.com/ssloxford/gsextract
6 https://www.wireshark.org

b) Custom Text Crawler

The second method comprises a custom text crawler that extracts string segments consisting of alphabet letters, numerals, and other ASCII characters. This script matches strings in the recorded byte-files with keywords from a list and collects the coinciding strings in a text file. This helped to identify satellite streams that were used for aviation-related communication.

The list of keywords (55 in total) was constructed to contain aviation-related search terms such as ‘air’, ‘aero’, ‘flight’, as well as other communication-related words like ‘wifi’, ‘connect’, and ‘update’. This list was used to automatically obtain relevant satellite feeds, where deeper manual analysis could then be further conducted. This method is more universal than GSExtract and can be employed on all DVB-S2 streams.

D. Ethical and Legal Considerations

As the data collected in this experiment comes from real-world network traffic, all recordings were treated with special care to adhere to the current legal regulations of the local jurisdiction.

The recordings and all follow-up data used for the analysis were stored on a secured server and were fully removed once they were no longer needed. The content of all messages was treated as if it contained sensitive data, as no knowledge about the sensitivity of the data existed in advance. Where applicable, affected parties have been informed directly.

4. FINDINGS

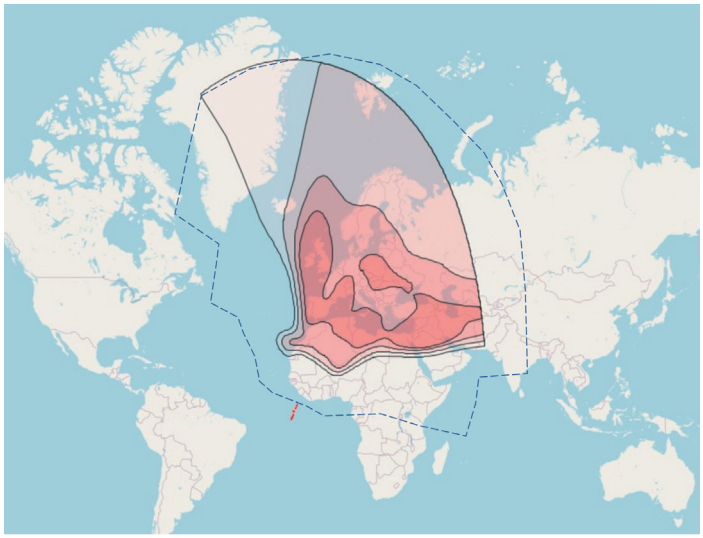
A total of 18 satellites were identified broadcasting in the Ku-Band frequencies as seen from our vantage point in Central Europe (Table II).

TABLE II: INFORMATION ON THE IDENTIFIED SATELLITES AND THEIR OBSERVED BEAMS. AS THEY ARE STATIONED IN GEOSTATIONARY ORBIT ABOVE THE EQUATOR, ONLY THE LONGITUDINAL COORDINATES ARE LISTED

Satellite Location	Satellite Name	Beam
1° West	Intelsat 10-02	Spot01
1° West	Thor 6	K2
1° West	Thor 5	T2
12° West	Eutelsat 12 West B	Europe
14° West	Express AM8	EuropeME
15° West	Telstar 12V	EuropeME
18° West	Intelsat 37e	Spot02
22° West	SES 4	EuropeME
24° West	Intelsat 905	Spot01
30° West	Hispasat 30W-6	EuropeNA
30° West	Hispasat 30W-5	Europe1E
5° East	Astra 4A	EuropeFSS
7° East	Eutelsat 7B	EuropeA
7° East	Eutelsat 7C	West
23° East	Astra 3B	PanEuropean
28° East	Astra 2E	Europe
28° East	Astra 2G	Europe
28° East	Astra 2F	Europe

We estimated their coverage footprint using the collected communications data, as shown by the dashed line in Figure 6. While each satellite footprint covers a different area, their complete collective footprint stretches around Europe and covers parts of the Atlantic Ocean, Northern Africa, and the Middle East.

FIGURE 6: THE COVERAGE FOOTPRINT OF THE MAIN AVIATION-RELATED SATELLITES OBSERVED IN OUR STUDY, REPRESENTED BY RED SHADES (LIGHTER SHADE ILLUSTRATING WEAKER SIGNALS). THE DASHED LINE INDICATES THE MAXIMUM OBSERVABLE AREA COVERED BY AT LEAST ONE SATELLITE LISTED IN TABLE II



Scanning these satellite positions for transmissions identified 34 frequencies, 26 from satellites positioned in the West and eight from satellites in the East. Over the course of 25 days in 2021, a sample file of 500 MB was recorded for each frequency.

The text crawler then identified five frequency recordings with potential aviation-related content. Since it was unknown from which satellite the signal of these beams was originating, a new set of longer recordings was conducted for all five constellations (Table III).

TABLE III: LIST OF DETAILED RECORDINGS OF BEAMS WITH AVIATION-RELATED CONTENT

Satellite Stream			Recording		
Location	**Frequency**	**Symbol Rate**	**Nr.**	**File Size**	**Duration**
12° West	11106 MHz	46657 KS/s	1	25.5 GB	73.5h
14° West	10984 MHz	51418 KS/s	2	67.0 GB	25.3h
12° West	11106 MHz	46657 KS/s	3	73.7 GB	20.5h
15° West	10985 MHz	51419 KS/s	4	87.7 GB	32.2h
15° West	11106 MHz	46657 KS/s	5	116.0 GB	28.3h

From inspecting the content of the newly recorded files, it turned out that the transmissions were sent from a high-throughput satellite, one of the 'leading satellites for mobile broadband maritime and aero services' [18]. Due to an incompatibility with the current version of GSExtract, the content of this feed was analysed exclusively using the text-crawling method. The red-shaded areas in Figure 6 show the footprint of these aviation-related satellite transmissions, with the lighter shades depicting weaker signals. In the following, we present first some results from an exploratory approach, followed by a more systematic analysis.

A. Exploratory Findings

Using an exploratory analysis, we present three main findings with regard to aviation-related communication on SATCOM streams from our study.

a) In-Flight Entertainment and Live Television

First, we observed unencrypted data originating from one of the world's leading service providers for in-flight entertainment (IFE) systems, which enables on-board live coverage of sports events, news broadcasts and other television programs with its global network of high-throughput satellites.

The messages contained technical information about the TV stream, such as encoding, aspect ratio, resolution, and audio language. It was also possible to obtain the port numbers of the channels and the private IPv4 address used for on-board transmission (Figure 7). During the period of the study, we observed 17 different TV channels, from international news to sports and entertainment.

FIGURE 7: MESSAGE PACKET CONTAINING INFORMATION ABOUT A CHANNEL PROVIDED BY A LEADING IFE SYSTEM

udioVideo","callSign":"NHK","image":"NHK","description":"Japanese-language chann
el designed to inform and entertain Japanese living and travelling outside Japan
. ","active":"1","name":W
"NHK World Premium","GndStreamAddress":"[redacted]","GndStreamPort":"[redacted]","v
ideo":{"id":"9","pid":"2048","resolution":"352x480","encoding":"h.264","aspect":
"16:9"},"audio":[{"id":"
12","pid":"8001","language":"jpn"}],"groups":[{"id":"1054"},{"id":"1001"},{"id":

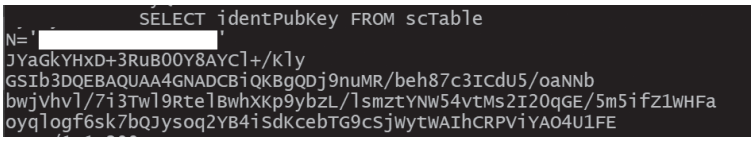
Knowing the IP addresses and ports from which these broadcasts are streamed may allow an attacker to hijack the television transmission by packet spoofing, imitating the stream's origin and replacing the content with their own. As an aside, as the captured messages did not indicate any form of encryption or authentication, this attack could become a real threat from someone inside the on-board network.

b) SQL Queries

Aside from data stemming from on-board entertainment systems, all five extensive recordings contained messages with two particularly noteworthy types of SQL

queries. The first one retrieves what looks like a public key from a database (Figure 8), opening up potential man-in-the-middle attack vectors, where public keys are replaced with malicious ones in transit. This makes it possible to decrypt and relay communication intended for the holder of the compromised public key.

FIGURE 8: OBSERVED SQL QUERY THAT RETRIEVES A PUBLIC KEY FROM A DATABASE



A second regularly observed and potentially vulnerable query inserts a commit into a database (Figure 9). Aside from a timestamp indicating the entry time of the log, different IDs and a 16-digit smart-card serial number are logged into the database. Smart cards can come in different shapes and are used for a wide range of purposes, from access control or authentication to financial transactions with a credit card. Having these queries sent openly and unauthenticated introduces, for example, the risk of replay attacks, where an attacker could capture the messages and send them again, possibly altering their content, such as the timestamp in the process.

FIGURE 9: EXAMPLE QUERY LOG COMMISSIONING SMART-CARD ENTRIES WITH TIMESTAMPS AND IDS

c) RSA Encryption Keys

Finally, the analysed transmissions contained large parts of private RSA keys in the PKCS#1 standard[7] (Figure 10). While the keys were not necessarily complete (due to lossy transmissions), RSA private keys should naturally never be exposed – even in part – due to their sensitive nature. The holder of a compromised private key should revoke and replace it immediately, as it opens up trivial attacks on the confidentiality and authenticity of any communication encrypted or signed with this private key.

Along with the private keys, the feeds also contained much shorter RSA public keys, which were complete and intact. This may pose a problem in specific man-in-the-middle circumstances, as discussed in the previous section.

7 https://datatracker.ietf.org/doc/html/rfc8017

FIGURE 10: AN INCOMPLETE (AND REDACTED) RSA PRIVATE KEY FOUND IN THE TRANSMISSIONS

B. Systematic Analysis of Aircraft and Stakeholders

We now analyse the type and owners of the aircraft seen during the study. In total, we could identify 328 different aircraft across 22 operators based on their tail numbers (see Table IV). These numbers were part of a message type captured in the recordings. Tail numbers are sufficient to identify an aircraft in conjunction with freely available online databases that reveal the aircraft type, its carrier airline, ICAO hex-codes and operator history [2].

TABLE IV: OVERVIEW OF AIRCRAFT IDENTIFIED BY THEIR BROADCASTED TAIL NUMBER

# of Aircraft	Operators	Registration Prefix	Registration Country
1	Japan – Air Self Defence Forces	-	Japan
11	Singapore Airlines	9V	Singapore
73	Emirates, Etihad Airways	T2	UAE
11	China Eastern Airlines, Air China, Cathay Pacific	B	China
37	Lufthansa	D	Germany
3	Iberia, Air Europe	EC	Spain
1	Alitalia	EI	Ireland
9	Air France, Aeroflot	F	France
7	British Airways, Virgin Atlantic	G	UK
2	Swiss	HB	Switzerland
2	Japan Airlines	JA	Japan
82	United Airlines, American Airlines, Aeromexico	N	USA
18	KLM	PH	Netherlands

3	Middle East Airlines	T7	San Marino
68	Turkish Airlines	TC	Turkey
328 aircraft	**22 airlines**		

Examples include aircraft used exclusively by the Japanese Prime Minister or members of the Imperial family for international travel (a Boeing 777-300ER). At the time, it was flying back from the 2021 G7 summit, which took place in Cornwall, UK, from 11 to 13 June.

Other notable aircraft identified were an Airbus A330-243 Prestige belonging to the Turkish Government and a Boeing 737-800 BBJ2 used for presidential flights of the United Arab Emirates. All three aircraft are blocked on flight tracking sites such as Flightradar24,[8] indicating a desire for privacy by their operators. As the captured messages circumvent these blocks by revealing concrete flight activity, we can see that even the most sensitive stakeholders can be affected by the lack of security on SATCOM links.

Our deeper analysis indicates that most aircraft using SATCOM are modern, wide-body aircraft belonging to major national flag carriers. This intuitively makes sense, as these airlines are generally more prone to invest in SATCOM connectivity and offer access to advanced entertainment systems, such as on-board Wi-Fi or live television, than low-cost carriers. The aircraft types found in the recordings also support this, as the benefit of extensive entertainment systems is assumed to be far greater on wide-body aircraft that cover long distances than on smaller, short-haul ones.

5. DISCUSSION

We now discuss our findings from the study. Table V presents an overview of the different issues found in our study and their potential impact on safety and security for passengers, crew, and operators.

8 https://www.flightradar24.com

TABLE V: OVERVIEW OF OBSERVED SATCOM ISSUES AND THEIR IMPACT ON SAFETY AND PRIVACY, ALONG WITH POTENTIAL ATTACK SCENARIOS EXPLOITING THE VULNERABILITY

Discovered Issue	Safety Impact	Privacy Impact	Attack Scenario
Wi-Fi Login Sites Visible	None	Potential for Strong Impact	Website Phishing
IFE system vulnerable	Potential for medium impact	None	Content spoofing
RSA private key messages	Impact unknown	Impact unknown	Message decryption
SQL queries visible	Impact unknown	Impact unknown	Replay attack
Tail numbers visible	Minor impact	Minor impact	Flight-path tracking

A. Communication Content Types

Our results show that widely used IFE and on-board Wi-Fi systems make use of unencrypted SATCOM channels for data transmission, allowing access to non-end-to-end-encrypted (E2EE) communication. On a cautiously positive note, and different to the maritime domain studied in [3], no email messages containing personal information from passengers were found during our time-limited study. A potential reason for this could be the use of E2EE by the observed systems. However, based on the recent literature, many non-E2EE systems are still in use, and examples may yet be found if the duration and scope of the study were extended.

As our study indicates, the use of insecure SATCOM systems on aircraft goes far beyond these passenger-oriented services. Traces of SQL queries and commits concerning public key and smart-card infrastructure are clear giveaways of crew- and business-related usage and several potential attack vectors.

As smart cards can be used for a large variety of tasks, it becomes more challenging to narrow down the purpose of the cards observed in our particular case. However, it seems reasonable to infer that the logs are used in the context of an employee time management system.

Other communication was not as readily identifiable using the applied methodology. However, we can speculate that observed XML files containing additional information about the carrier airline and a version number are likely part of the ‘Aircraft Earth Station (AES)’ management. AES comprises the setup and the billing of the SATCOM provider, providing a potentially crucial entry point for impersonation attacks.

B. Privacy Implications and Safety Concerns

In this part, the safety and privacy impacts of the findings are put into perspective with the likelihood of an attack exploiting these issues.

Starting with the general observation that the studied SATCOM channels do not deploy default encryption, simple eavesdropping attacks may result in a substantial breach of privacy. Although no emails or other personal messages were found during the present study, passengers of an airline offering unencrypted on-board internet access via a satellite connection are likely to leak private information over time. This may, for example, happen to unsuspecting users downloading their emails in clear text from an unsecured mail-server using POP3, as shown previously in the maritime context [3]. Artefacts in related studies also revealed mobile phone traffic originating from aircraft. [13]

The ability to track aircraft based on their SATCOM connections provides a potential operational security risk, as outlined in detail in previous literature (e.g. [2]). While not a concern for most commercial airlines, as their flight paths are well known and easily accessible, some of the identified aircraft actively hide their whereabouts from the public. All three observed government airplanes were either fully or partly blocked on public tracking sites, underlining their individual desire for privacy.

Aside from potential privacy concerns for crew and passengers, our study did not directly indicate message content, which could pose a direct risk for the safety of an aircraft. In particular, no connection to the safety-critical aircraft control domain was observed, which would make it possible to tamper with flight control systems.

However, the transmissions directed at the in-flight entertainment system could still have a direct impact on on-board safety. As previously described, message content for live television services contained information about which ports and local IP addresses were used to stream content to seat monitors or private devices. An attacker on board the aircraft and connected to the local network could try to perform an IP spoofing attack, imitating the stream origin, and replacing the television transmission with their own. Depending on the content of this new stream, this could lead to disinformation or unrest among the passengers, an attack vector suggested previously by Ruben Santamarta of IOActive [14].

Another potential safety compromising attack could make use of the clear-text database accesses. While the observed SQL queries do not seem to contain content concerning safety or privacy, the database itself could become the target of a directed attack. A malicious adversary could try to access the entire content stored in the database, which might result in a severe leak of private information. Alternatively, he could

carry out a replay attack of already seen queries or forge completely new queries. Although the exact purpose of the database and the logging queries are not clear, it is more than reasonable to assume that such interference would strongly impact any system working with it.

The last finding with the potential to affect safety or privacy involves transmitted RSA private keys. As it was not possible to identify why or to whom they were sent, it is difficult to measure their direct impact. Nonetheless, the application or protocol that sends these messages is strongly violating any good practice in information security, and further investigations may lead to identifying further vulnerabilities.

C. Countermeasures

Finally, we review possible solutions to the issues discussed with respect to the specific environment of the aviation industry. There are two principal levels at which improvements to the safety and privacy of satellite communication can be applied.

The first option involves the satellite service providers, who are in the best position to improve the confidentiality of any data sent over their satellite network. As previous studies in other domains have shown, the industry-standard level of protecting satellite communication is insufficient and should be strongly reconsidered [3]. In response to this non-satisfactory situation, the academic research community has recently studied protocols and methods for provable secure satellite communication systems (e.g. [15], [16]). One of these promising approaches is QPEP, a protocol built on top of the QUIC standard, providing the performance benefits of industry-standard Performance Enhancing Proxies (PEPs) while offering the security of end-to-end message encryption at the same time [17].

As satellite service providers might not be inclined or able to make these changes to the data security of their satellite network traffic quickly (e.g. due to operational or financial reasons), another approach could see those manufacturers responsible for the development of IFE and on-board Wi-Fi systems improve their security practices. In addition to reviewing the use of SATCOM for sensitive content, this could involve the utilization of higher layer end-to-end encryption for all applications available on board.

However, with the aviation industry's historically strong focus on the safety of systems with multiple redundant layers, providing security guarantees on this level could be very challenging. A research project by Bernsmed *et al.* [10] on this topic advocates for a system where both sides provide certain security guarantees, stating that: 'SATCOM datalink systems must enable integrity protection and data-origin authentication of the

datalink applications, whereas confidentiality and non-repudiation protection should be implemented on an application-by-application basis.' [10]

Aside from any technical countermeasures, raising awareness of the issues with SATCOM usage in the aviation domain will be pivotal for the successful implementation of safe and secure datalink communication systems in the future.

D. Limitations

As our study was carried out using low-end commercial off-the-shelf equipment accessible to unsophisticated threat actors, some natural limitations exist. While the quality of the recorded files was sufficient to explore aviation-related content, the transmitted data was often cut short or corrupted due to lossy connections during the recording. It also needs to be stated that this experimental setup could only intercept satellite downlink messages sent towards an aircraft. To analyse different satellite streams, the receiver station would need to be positioned differently – in proximity to a ground station.

E. Further Research

The insights gained from this project represent interesting new options for future research.

One such topic could be the identified possibility of spoofing attacks on the in-flight entertainment system. Follow-up research could investigate the feasibility of such an attack, which might lead to a general study on the safety and security of on-board wireless networks.

Other studies could focus on the origin of the observed messages – for example, looking into which protocols or algorithms sent the RSA keys or the aircraft tail numbers. For these projects, it would prove beneficial to adapt the GSExtract tool to handle more relevant types of data recordings and collect detailed statistics on the type of network traffic similar to [3].

Finally, as this project only focused on satellite downstream transmissions, an interesting new approach would certainly be to intercept satellite upstream messages sent by an aircraft towards a satellite. Even though this requires an entirely different setup, the results would provide new insights into aircraft-to-ground satellite communication.

6. SUMMARY AND CONCLUSION

The goal of this work was to find and analyse safety and privacy-related issues of satellite communications inside the aviation domain. The findings from the conducted experiment and the subsequent analysis of the recorded files prove the existence of such problems and show that an eavesdropper can intercept SATCOM messages using widely available low-budget equipment. The results included streaming data related to the in-flight entertainment system of modern aircraft, as well as different messages containing SQL queries, on-board Wi-Fi addresses and RSA private keys.

With the present study, we want to raise awareness regarding the fact that the current state of play may compromise the confidentiality and integrity of sensitive aviation data, as attackers may try to exploit them. We believe that stakeholders inside the aviation industry need to adapt to the new threat environment of cheap and easily accessible satellite communications. As the future of air traffic management is going to rely heavily on SATCOM-based communication, security concerns need to be addressed with the same priority as safety is currently. As shown by previous works, the interconnected air traffic management systems of the future cannot be deemed safe while their security is not guaranteed. [8]

In addition to presenting technical approaches to secure satellite communication channels, this research has also illustrated the importance of raising public awareness of this topic, as all stakeholders travelling on SATCOM-connected aircraft can also be affected without their knowledge.

REFERENCES

[1] ICAO, 'The World of Air Transport in 2019', 2019. Accessed: Aug. 13, 2021. [Online]. Available: https://www.icao.int/annual-report-2019/Pages/the-world-of-air-transport-in-2019.aspx

[2] M. Strohmeier, D. Moser, V. Lenders, M. Smith, M. Schäfer, and I. Martinovic, 'Utilizing Air Traffic Communications for OSINT on State and Government Aircraft', in *Cyber Conflict (CyCon) 2018 10th International Conference*, Tallinn, Estonia, 2018.

[3] J. Pavur, D. Moser, M. Strohmeier, V. Lenders, and I. Martinovic, 'A Tale of Sea and Sky: On the Security of Maritime VSAT Communications', in *2020 IEEE Symposium on Security and Privacy (S&P)*, 2020.

[4] Panasonic Avionics Corporation, 'Why Passengers Are Demanding Live Television', 2020. Accessed: Aug. 3, 2021. [Online]. Available: https://www.panasonic.aero/our-offerings/solutions/theatre/live-television/#download-book

[5] K. Sampigethaya, R. Poovendran, S. Shetty, T. Davis, and C. Royalty, 'Future E-Enabled Aircraft Communications and Security: The Next 20 Years and Beyond', in *Proceedings of the IEEE*, Nov. 2011, vol. 99, pp. 2040–2055.

[6] M. Smith, D. Moser, M. Strohmeier, V. Lenders, and I. Martinovic, 'Undermining Privacy in the Aircraft Communications Addressing and Reporting System (ACARS)', in *Proceedings on Privacy Enhancing Technologies*, Jun. 2018, vol. 2018, pp. 105–122.

[7] M. Strohmeier, I. Martinovic, and V. Lenders, 'Securing the air–ground link in aviation', in *The Security of Critical Infrastructures*. Cham, Switzerland: Springer, 2020, pp. 131–154.

[8] M. Smith, M. Strohmeier, J. Harman, V. Lenders, and I. Martinovic, 'A View from the Cockpit: Exploring Pilot Reactions to Attacks on Avionic Systems', in *The Network and Distributed System Security Symposium* (NDSS), 2020.
[9] M. Strohmeier, A. K. Niedbala, M. Schäfer, V. Lenders, and I. Martinovic, 'Surveying Aviation Professionals on the Security of the Air Traffic Control System', in *International Workshop on Cyber Security for Intelligent Transportation Systems*, 2018, pp. 135–152.
[10] K. Bernsmed, C. Frøystad, P. H. Meland, and T. A. Myrvoll, 'Security requirements for SATCOM datalink systems for future air traffic management', in *2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)*, 2017, pp. 1–10.
[11] R. Santamarta, 'Last Call for SATCOM Security', IOActive, 2018. [Online]. Available: https://ioactive.com/wp-content/uploads/2018/08/us-18-Santamarta-Last-Call-For-Satcom-Security-wp.pdf
[12] M. Strohmeier, D. Moser, M. Schäfer, V. Lenders, and I. Martinovic, 'On the Applicability of Satellite-Based Air Traffic Control Communication for Security', *IEEE Communications Magazine*, vol. 57, no. 9, pp. 79–85, 2019.
[13] J. Pavur, 'Whispers Among the Stars', in *DEF CON "Whispers Among the Stars: A Practical Look at Perpetrating (and Preventing) Satellite Eavesdropping Attacks." In Conference briefing. Conference briefing. Black Hat USA*. Las Vegas, NV, Aug, vol. 5. 2020 Safe Mode, 2020.
[14] R. Santamarta, 'In Flight Hacking System', IOActive, Dec. 12, 2016. [Online]. Available: https://ioactive.com/in-flight-hacking-system/
[15] K. Guo, K. An, B. Zhang, Y. Huang, X. Tang, G. Zheng, and T. A. Tsiftsis, 'Physical Layer Security for Multiuser Satellite Communication Systems with Threshold-Based Scheduling Scheme', *IEEE Transactions on Vehicular Technology*, vol. 69, pp. 5129–5141, 2020.
[16] M. Qi and J. Chen, 'An enhanced authentication with key agreement scheme for satellite communication systems', *International Journal of Satellite Communications and Networking*, vol. 36, pp. 296–304, 2018.
[17] J. Pavur, M. Strohmeier, V. Lenders, and I. Martinovic, 'QPEP: An Actionable Approach to Secure and Performant Broadband from Geostationary Orbit', The Network and Distributed System Security Symposium (NDSS), 2021.
[18] Telesat, 'Telstar 12 VANTAGE', Dec. 2020. [Online]. Available: https://www.telesat.com/wp-content/uploads/2020/12/Telstar-12-VANTAGE.pdf

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

Keep the Moving Vehicle Secure: Context-Aware Intrusion Detection System for In-Vehicle CAN Bus Security

Sampath Rajapaksha
School of Computing
Robert Gordon University
Aberdeen, United Kingdom
s.rajapaksha@rgu.ac.uk

Harsha Kalutarage
School of Computing
Robert Gordon University
Aberdeen, United Kingdom
h.kalutarage@rgu.ac.uk

M. Omar Al-Kadri
School of Computing and Digital Technologies
Birmingham City University
Birmingham, United Kingdom
Omar.alkadri@bcu.ac.uk

Garikayi Madzudzo
Horiba-MIRA
Coventry, United Kingdom
Garikayi.madzudzo@horiba-mira.com

Andrei V. Petrovski
School of Computing
Robert Gordon University
Aberdeen, United Kingdom
a.petrovski@rgu.ac.uk

Abstract: The growth of information technologies has driven the development of the transportation sector, including connected and autonomous vehicles. Due to its communication capabilities, the controller area network (CAN) is the most widely used in-vehicle communication protocol. However, CAN lacks suitable security mechanisms such as message authentication and encryption. This makes the CAN bus vulnerable to numerous cyberattacks. Not only are these attacks a threat to information security and privacy, but they can also directly affect the safety of drivers, passengers and the surrounding environment of the moving vehicles. This paper presents CAN-CID, a context-aware intrusion detection system (IDS) to detect cyberattacks on the CAN bus, which would be suitable for deployment in automobiles, including military

vehicles, passenger cars and commercial vehicles, and other CAN-based applications such as aerospace, industrial automation and medical equipment. CAN-CID is an ensemble model of a gated recurrent unit (GRU) network and a time-based model. A GRU algorithm works by learning to predict the centre ID of a CAN ID sequence, and ID-based probabilistic thresholds are used to identify anomalous IDs, whereas the time-based model identifies anomalous IDs using time-based thresholds. The number of anomalies compared to the total number of IDs over an observation window is used to classify the window status as anomalous or benign. The proposed model uses only benign data for training and threshold estimation, avoiding the need to collect realistic attack data to train the algorithm. The performance of the CAN-CID model was tested against three datasets over a range of 16 attacks, including fabrication and more sophisticated masquerade attacks. The CAN-CID model achieved an F1-Score of over 99% for 13 of those attacks and outperformed benchmark models from the literature for all attacks, with near real-time detection latency.

Keywords: *controller area network, anomaly detection, vehicle networks, CAN bus*

1. INTRODUCTION

Modern automobiles are becoming complex and highly connected to provide safe, efficient and intelligent services to users. To facilitate these services, automobiles are equipped with multiple networks and communication devices and a range of sensors, actuators, cameras and microprocessor-based electronic control units (ECUs) [1]. Modern vehicles run software that exceed 100 million lines of code, and future vehicles will require 200 to 300 million lines of code [2]. These software run on up to 100 ECUs [3] that are connected to a controller area network (CAN) which is considered to be the de-facto network protocol for in-vehicle communication [1]. The CAN bus is a message-based protocol commonly used in vehicles, aerospace, industrial automation and medical equipment due to several benefits such as being low cost, speedy, lightweight and robust [4]. Despite these benefits, the CAN bus lacks security measures, especially given the absence of authentication, an ID-based priority system, broadcast transmission and lack of encryption. Increased connectivity and complexity and CAN bus security flaws have made modern vehicles vulnerable to cyberattacks. In fact, security researchers have demonstrated the capability of attacks against modern vehicles by compromising the CAN networks of various vehicle brands [5]–[7]. These researchers have shown that it is possible to implement CAN message injection attacks remotely and take physical control of these vehicles. An attacker obtaining physical control of a moving vehicle will directly affect the safety

of drivers, passengers and the surrounding environment of the vehicle. The security of modern automobiles is a major concern for automotive manufacturers; therefore, they are seeking security measures to protect against such attacks [8], [9].

Developing an in-vehicle IDS for widespread adoption with a high detection capability is challenging due to the lack of knowledge about the CAN data specifications [10]. Generally, specifications of CAN messages are stored in a database-like file known as the database CAN (DBC), a confidential source of proprietary information, access to which is usually restricted to the vehicle manufacturer. Depending on the number of ECUs, the CAN bus transmits about 2000 frames per second [11]. This demands an IDS with real-time or near real-time detection capability under a computationally constrained environment. Cyberattackers could use various types of attacks (e.g. injection and masquerade attacks) that alter the different data fields of CAN messages to compromise the in-vehicle network. This is another challenge that limits the detection and generalization capabilities of an IDS. In addition, many events that arise in a vehicle could be considered anomalies despite being legitimate driving scenarios. For example, an emergency brake or sudden steering wheel turn while driving at 70 mph would be considered anomalous in normal driving scenarios. These kinds of benign anomalous behaviours could produce a significant number of false positives. Hence, knowledge of the context of the CAN sequences is vital to distinguish benign anomalies from potential attack scenarios. To successfully deal with the aforementioned challenges, this paper proposes CAN-CID (CAN Centre ID prediction), a novel context-aware ensemble IDS for the CAN bus based on natural language processing (NLP) and time-based techniques.

The main contributions of this paper can be follows.

1. CAN-CID uses only benign data to train the model and estimate thresholds. This avoids the need to collect real attack data to train the algorithm. It is significantly easier and safer to collect benign CAN data from real vehicles than to collect attack data. Further, using only benign data (one-class) during the training process improves the generalization capability of the algorithm.
2. Probability-based thresholds were estimated for each ID using only benign training data. Minimum thresholds were selected with the aim to minimize false positives, which will help to improve the overall accuracy of the ensemble model.
3. CAN-CID uses a one-layer shallow GRU network to detect anomalous ID sequences. Hence it is lightweight, and detection latency is very low (10 ms for a 100 ms window). This makes the proposed solution suitable to deploy in real vehicles.

The rest of this paper is structured as follows: Section 2 presents the related work. Section 3 provides the background, including CAN data analysis. The proposed algorithm is explained in section 4. In section 5, the experiment results and performance evaluations are presented. Finally, section 6 concludes the paper.

2. RELATED WORKS

Recent experiments focusing on attacks against modern automobiles [5], [6] have motivated research into countermeasures against in-vehicle network attacks. The majority of these works have focused on securing the CAN bus, as notable experimental attacks targeted the vulnerabilities of the CAN bus [6], [7]. In [11], the authors proposed a specification-based IDS for in-vehicle network intrusion detection by extracting design specifications of CAN messages. The IDS proposed in [12] used unique voltage signals generated by ECUs as features of the deep support vector domain description model. However, both of these models [11], [12] have a low generalization capability as they require specific knowledge of CAN data. A one-class compound classifier was used in [13] to detect CAN bus attacks. But this detected only 45–65% of attacks. The authors then suggested an ensemble of detection methods to overcome the problems that arise when using only one classifier. In [14], the authors proposed a long short-term memory (LSTM) autoencoder to detect CAN bus anomalies. This was trained using a payload of legitimate CAN frames. Reconstruction error was used to distinguish benign from malicious frames. A major limitation of this model, however, was the slow computation time due to the complex model architecture. LSTM-based deep learning model, which utilized the linear embedding of the CAN payload, was used in [15] to detect contextual anomalies in the payload. The authors examined the effect of context by removing the embedding layer of the proposed model. They observed that context and embedding helped to slightly improve the performance. However, this model recorded only around 95% accuracy for all attacks on one dataset. CAN payload signals were selected as the features of the deep learning-based IDS proposed in [16]. Similarly, the authors in [17] used sensor values in their deep neural network-based IDS. However, both of these approaches [16], [17] require the DBC files or knowledge about the CAN payload, which could limit the generalization capability of the proposed algorithms.

Frequency or time-based IDSs utilize the timing of CAN frames or the sequential nature of the IDs. In [18], the authors developed a context-aware anomaly detector for monitoring cyberattacks on the CAN bus using sequence modelling. The authors of [19] proposed an anomaly detection algorithm by modelling the normal behaviour of the CAN bus considering the recurring pattern of CAN IDs. This is equivalent to 2-grams in the N-gram-based model used in [18]. While N-gram-based algorithms can

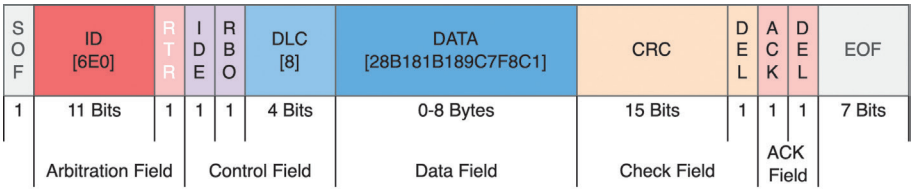
capture the context, this often leads to high computational overhead as N increases. A time-based IDS was proposed by [20] to detect CAN injection attacks. In [21], the authors used an LSTM model to predict the next ID and compare it with the actual ID to identify anomalous frames. However, this approach achieved only 60% accuracy. A CAN bus attack detection framework was proposed by [22] utilizing both a rule-based and a supervised LSTM model. This ensemble model outperformed the individual models. In general, the deep learning-based IDSs discussed above demonstrated a higher detection capability than the other models. However, supervised learning-based models might have low generalization capability to other attacks and vehicles as they learn the attack pattern of the particular dataset. Further, these models have high detection latency due to their complex deep learning architecture. To address these problems, this work presents a lightweight ensemble model that uses a shallow neural network.

3. BACKGROUND

A. Controller Area Network (CAN Bus)

CAN is a broadcast-based communication protocol developed by Bosch for in-vehicle communication [23]. ECUs of modern vehicles communicate using high-speed and low-speed CAN buses as their network protocol. Time-critical modules such as engine control and transmission control are connected to a high-speed CAN bus, whereas less time-critical modules such as door control and light control are connected to a low-speed CAN bus. A CAN bus data frame includes several fields: the CAN ID (arbitration field) is used to prioritize the messages and is capable of handling concurrent messages; a CAN payload contains the actual information (data) that is to be transmitted over the network; and other fields include start of frame (SOF), control field (DLC), cyclic redundancy code (CRC), acknowledge field (ACK), and end of frame (EOF). These fields are depicted in Figure 1, with their respective bit-lengths. When a node (ECU) is ready to transmit a frame, it checks the status of the bus, and if the bus is idle, it transmits the frame. The addresses of the transmitting node and the receiving node are not included in the frame. Instead, it uses CAN IDs unique to the transmitting nodes. As a result of the broadcast nature of the network, all nodes in the CAN network can receive the frame. Based on the ID of the frame, other nodes in the network decide to accept or ignore the frame. The priority-based arbitration scheme ensures that the highest priority IDs (lower IDs) get bus access when multiple nodes simultaneously transmit frames onto the CAN bus. The lowest priority IDs, on the other hand, must wait until the bus becomes idle.

FIGURE 1: CAN BUS DATA FRAME. EXAMPLE VALUES ARE GIVEN FOR ID, DLC AND DATA FIELDS



B. CAN Bus Vulnerabilities and Attacks

The CAN bus is designed to provide robust, efficient, simple and low-cost in-vehicle communication without paying much attention to security-related features [13]. Therefore, by design, it is vulnerable to cyberattacks. Since the CAN bus uses no authentication, any node could transmit a message with an ID that belongs to another node. In addition, CAN frames are not encrypted due to real-time communication requirements. This allows attackers to collect and analyse CAN data (via sniffing). The broadcast nature of the CAN bus also transmits frames to all nodes connected to the CAN bus. Therefore, by utilizing this property, a compromised ECU can not only monitor and listen to all CAN frames transmitted through the CAN bus but also send any frame to the network. Furthermore, attackers can use the ID-based priority scheme to inject their messages with the highest priority IDs to create a denial-of-service (DoS) attack, consequently making communication services unavailable to other IDs.

Some of the common CAN attack types are: DoS [24], fuzzing [25], replay [24], spoofing [26] and masquerade attacks [27]. In a fuzzing attack, a large number of random frames are injected into the CAN bus. Replay attacks re-send previously recorded frames at different times. When an attacker targets (injects) frames with specific CAN IDs, it is called a spoofing attack. In a masquerade attack, a compromised node impersonates another node to send malicious frames. All of these attacks have the potential to cause unexpected or harmful effects to a vehicle depending on the attacker's purpose.

C. CAN Bus Data Analysis

We analysed CAN ID data to understand the anomalous traffic patterns, both benign and malicious. To do this, we used a publicly available dataset, the Real ORNL Automotive Dynamometer (ROAD) CAN Intrusion dataset [10]. Figure 2 shows a five-second snapshot of a targeted ID attack. In this attack, the targeted ID is 0D0. What stands out in this figure is the periodic behaviour of the IDs. Each node transmits frames at a fixed interval, as observed in [6]. In the ROAD dataset, 104 out of 106 IDs exhibit similar behaviour. However, the injected ID causes a change in this pattern

during the targeted ID attack. This can be observed in the shaded area (attack period) for 0D0. This changes the fixed transmission interval of IDs compared to that of the period of normal driving. In addition, it could create new ID sequences resulting from new frames appearing in an unusual context. For example, it might introduce a new ID sequence, such as '6E0 0D0 0D0', which was not observed during normal driving conditions.

FIGURE 2: FRAME TRANSMISSION OF A TARGETED ID (0D0) ATTACK. THE SHADED AREA REPRESENTS THE ATTACK PERIOD. THIS REPRESENTS ONLY A SUBSET OF THE 106 CAN IDS

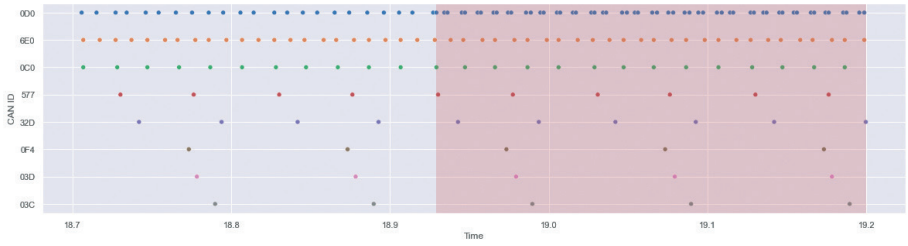


FIGURE 3: FRAME TRANSMISSION OF A MASQUERADE ATTACK (0D0). THE SHADED AREA REPRESENTS THE ATTACK PERIOD. THIS REPRESENTS ONLY A SUBSET OF THE 106 CAN IDS

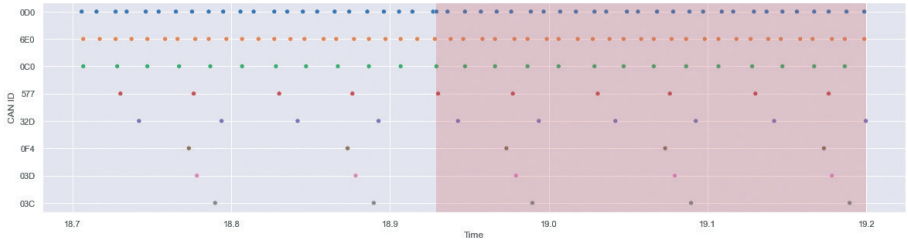


Figure 3 shows a five-second snapshot of a masquerade attack for the same ID (0D0). In this case, it does not significantly change the ID transmission frequency [10]. However, a masquerade attack might cause a slight deviation (shift of time) in the frame transmission time due to the difficulty of time synchronization with the legitimate ECU [28]. In addition, since a masquerade attack stops the frame transmission of a legitimate ECU, there might be a brief period where there is no frame transmitted with the targeted ID. A CAN bus transmits a large number of messages per second. Therefore, even a slight deviation from the normal driving scenario could create new ID sequences. For example, ID 0D0 might have the sequence 'ODO 6E0 0C0' during normal driving, whereas a slight time shift or absence of frames could create a new sequence of '6E0 0D0 0C0'. This behaviour (frequency and sequence change) can be observed for all injection and masquerade attacks for the ROAD dataset.

After analysing both benign and attack CAN traffic, our main finding is that most CAN IDs exhibit periodic behaviour that creates a finite set of ID sequences for a fixed window size (e.g. a window of ten consecutive IDs). Attacks on the CAN bus are likely to change the periodic behaviour of the IDs and hence create new sequences. In addition, injection attacks change the time between the consecutive attack IDs. Carefully trained machine learning algorithms can detect these subtle changes in CAN ID streams. These findings provide the basis for the proposed IDS.

4. PROPOSED CAN-CID MODEL

A. Threat Model and Datasets

In this work, we used the ROAD [10] dataset to test the proposed model. Additionally, to evaluate the generalization capability of the model, we used two other publicly available datasets, the car-hacking dataset for intrusion detection (HCRL CH) [29] and the survival analysis dataset for automobile IDS (HCRL SA) [30]. The ROAD dataset is considered the first open CAN bus dataset with advanced types of real attacks that have physically verified effects on the vehicle [10]. Data was collected through the OBD-II port in a fully compromised ECU mode while driving the vehicle on a dynamometer or on the road. The dataset includes 12 ambient (benign) datasets representing different driving activities, including drive, accelerate, decelerate, reverse, brake, cruise control, turn signals and anomalous but benign driving activities such as unbuckling a seatbelt and opening doors while driving. Attacks are categorized as fabrication attacks, suspension attacks and masquerade attacks. Fabrication attacks include fuzzing and targeted ID attacks. The attacks shown in Table I were selected to investigate the algorithm's detection capability.

TABLE I: HIGH-FREQUENCY INJECTION (FABRICATION) ATTACKS ON THE ROAD DATASET

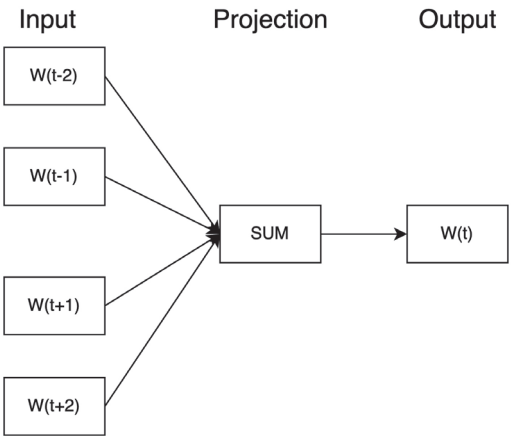
Attack	Attack technique	Consequence
Fuzzing	Inject random IDs and arbitrary payloads	Wide variety of unexpected results
Correlated signal attack	Inject false wheel speed values (ID-6E0)	Stop the car due to different pairwise wheel speeds
Max speedometer attack	Change one byte of payload to maximum (FF) value (ID-0D0)	Display false speedometer value
Reverse light on attack	Change one bit of payload (ID-0D0)	Reverse lights do not reflect what gear the car is using
Reverse light off attack	Change one bit of payload (ID-0D0)	Reverse lights do not reflect what gear the car is using

For fabrication attacks, the attacker injects a frame with a targeted ID immediately after a legitimate frame appears. The aim of this is to get the vehicle to ignore the legitimate message and accept the injected frame to change the vehicle state. In addition to the above attacks, this dataset includes a masquerade version of each ID fabrication attack. The masquerade version removes the legitimate target ID frames relevant to each injected frame in post-processing to simulate a masquerade attack. These realistic attacks required message injections for each attack. Such an approach was trialled by [6] in their experimental attacks. Message injections lead to changes in the transmission interval of the targeted ID and new sequences being created. Hence, even a payload attack might require an attacker to inject frames into the CAN bus [6]. This requirement makes an ID sequence or frequency-based model suitable for detecting the majority of payload attacks without using the payload-related features. The HCRL CH dataset includes DoS, fuzzy and spoofing (RPM and gear) attacks, whereas the HCRL SA (KIA Soul) dataset includes flooding, fuzzy and malfunction (targeted ID) attacks.

B. CAN Centre ID Prediction Task

This work is inspired by the work of [18] and the continuous bag-of-words (CBOW) model architecture proposed by [31]. In [18], the authors used N-gram distributions to build the CAN ID sequence model. The underlying concept of this work is a mathematical model (n-gram) that can be trained to learn the CAN message sequences and predict subsequent elements in the sequence. The authors showed that the occurrence of an event (ID) can be determined based on a short history. However, this might depend on the number of nodes in the network (equivalent to the number of unique words in a language). Thus for a larger number of nodes, a longer history may be required as a larger number of unique sequences could be created for the selected window size. N-gram models are inefficient for higher values of N because this will result in more combinations. This approach [18] is similar to the next word prediction task of NLP given the previous words (context). The Word2vec model proposed by [31] learns the word vectors (word embeddings) by learning to predict the centre (target) word given the context. This model architecture is shown in Figure 4. The CBOW model is expected to learn the word vectors representing the middle word's meaning and the context words. However, the main objective of the CBOW model is not to predict the words but to learn accurate word vectors that encode semantic relationships for all the words in the corpus. Then, the learned word vectors can be used in many language models with specific deep learning architectures.

FIGURE 4: CONTINUOUS BAG-OF-WORDS (CBOW) ARCHITECTURE TO PREDICT THE CENTRE WORD GIVEN THE PREVIOUS AND NEXT WORDS AS THE CONTEXT



Using the target word's historical and future words as the input words improved the centre word prediction [31]. We expect the same behaviour for CAN ID sequences. To elaborate on this, as an example, take a driving scenario of a right-hand turn at an intersection. Possible events in the vehicle are activate signal lights, decelerate (apply brake), stop, accelerate and turn right. If we want to predict the third event, which is stop in this case, we can use only previous events as the context or use previous and future events as the context. These two tasks can be formulated as follows:

$$x_1 = \{activate\ signal\ lights,\ decelerate\},\ y = \{stop\} \quad (1)$$

$$x_2 = \{activate\ signal\ lights,\ decelerate,\ accelerate,\ turn\ right\},\ y = \{stop\} \quad (2)$$

The second task (equation 2) can be used to make the prediction (stop) with higher accuracy, as the number of possible events for the centre (middle) event will be equal to or fewer compared to the first task. For example, 'accelerate' would be another probable prediction for equation 1. But in the second task, given accelerate in the context, it will make the prediction of 'stop' more accurate. Therefore, we use the CBOW architecture to infer the context for CAN ID sequences. One limitation of the CBOW approach is that it must wait for a few messages to see if the target (centre) ID is malicious. However, considering the CAN ID transmission rate, this will be a minimal amount of time (around a 0.005 s delay for 10 IDs). Additionally, continuous message injections are required to execute such an attack [6], which increases the

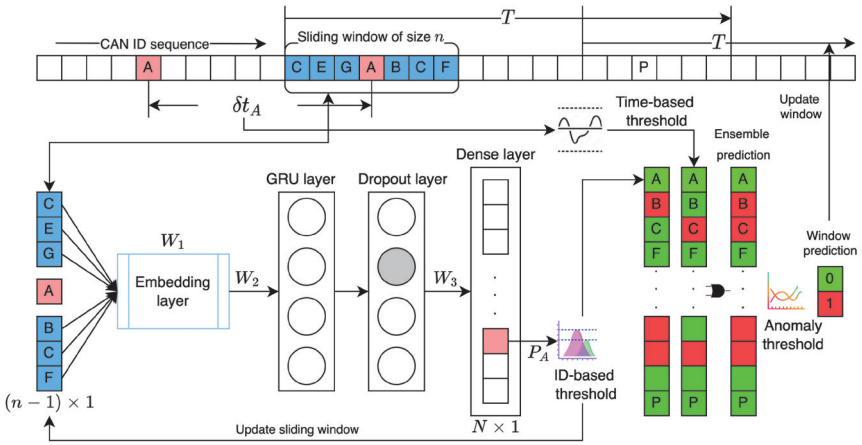
chances of detecting the attack before the attacker can take physical control of the vehicle. Hence, CBOW is a viable option for detecting attacks in CAN ID sequences.

C. CAN-CID Architecture

The order of the words is not considered in the CBOW model. However, the order is highly important to identify anomalous frames in CAN ID sequences. A recurrent neural network (RNN) model can capture the temporal patterns of sequential data [32]. But RNN models do not have a long-term memory due to the vanishing gradient problem. To address this issue, LSTM was introduced. LSTM consists of three gates: input, forget, and output. In contrast, GRU is a variant of LSTM with only two gates: reset and update. The simple structure reduces the matrix multiplication, making GRU more computationally efficient with low memory overhead [32]. Due to these properties, which are ideal for resource-constrained environments, we use a GRU layer to capture the temporal pattern of CAN ID sequences.

Figure 5 represents the architecture of the proposed model. We use a sliding window of size n (number of IDs) within a large sliding window of size T (time), where n is an odd number. Let N be the total number of unique IDs. For the GRU-based model, the input to the embedding layer is a sequence of vectorized CAN IDs of size $(n - 1)$. The centre (middle) ID $(n + 1)/2$ is used as the target of the prediction. As mentioned earlier, a single GRU layer is used as the hidden layer to learn the temporal patterns. A dropout layer is used to reduce the overfitting and improve the model generalization capability. The output layer is a dense layer that outputs softmax probabilities for N IDs. During the training, W_1, W_2, and W_3 are updated using backpropagation. P_A, which represents the softmax probability of the target ID (A), is compared with the pre-defined ID-based threshold. If the predicted probability is less than the threshold, the target ID is flagged as a weak anomaly; otherwise it is flagged as a benign ID. In the same way, the time-based model compares the time between two consecutive target IDs (δt_A) with the pre-defined time-based thresholds (minimum and maximum time). If δt_A is outside the thresholds, the current ID is flagged as a weak anomaly; otherwise it is flagged as a benign ID. This process continues for all IDs in window T. The OR operator is used to combine the two models as an ensemble model. Finally, an anomaly threshold is used to classify the window of time T as a malicious sequence or a benign sequence. This process continues for all IDs in the CAN ID stream by sliding the window of time T. The sliding window overlaps for $(n - 1)$ IDs, to make predictions for the missing IDs from the previous window.

FIGURE 5: ENSEMBLE MODEL ARCHITECTURE



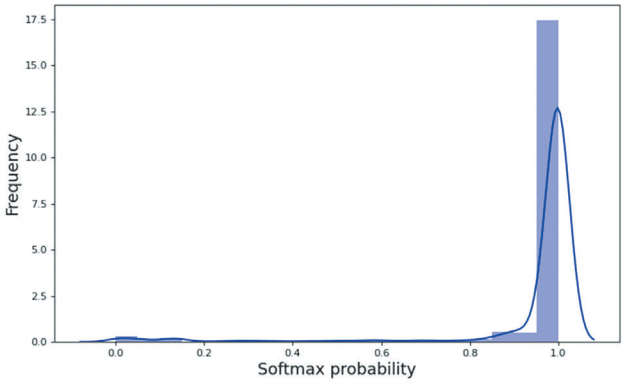
D. Threshold Estimation

The proposed model uses three thresholds.

1) ID-Based Threshold

A sample of the benign dataset was used to estimate thresholds. Softmax probabilities were calculated for all IDs in the benign sample. The minimum values of each ID were selected as the ID-based thresholds to minimize the false positives of the ensemble model. We assumed a zero probability of values less than the minimum values for benign data. Figure 6 shows a softmax probability distribution for a selected ID.

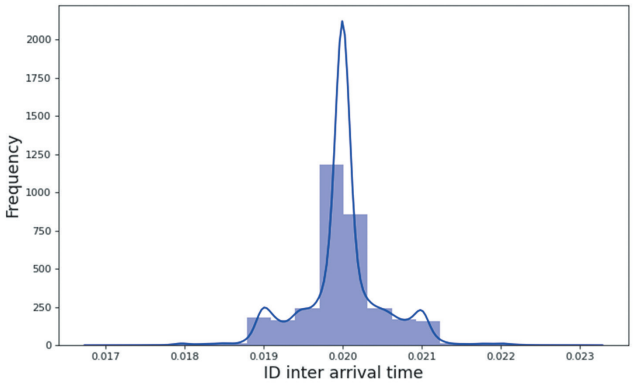
FIGURE 6: SOFTMAX PROBABILITY DISTRIBUTION OF ID 580



2) Time-Based Threshold

We used the training dataset to define the time-based threshold. For each ID, we calculated the time difference between two consecutive frames for the benign dataset. Then, the minimum and maximum values for each ID were used as the minimum and maximum thresholds. Figure 7 shows an inter-arrival time distribution for a selected ID.

FIGURE 7: INTER-ARRIVAL TIME DISTRIBUTION FOR ID 580



3) Anomaly Threshold

We used ID-based and time-based thresholds to identify weak anomalies. Counting weak anomalies over a window (T) helps to minimize false positives. Hence, we defined the anomaly thresholds (Ɛ) to identify the windows as attack or benign. We assigned labels for each window (0 for benign and 1 for attack) as the ground truth and used the same to evaluate the model performance. Equation 5 was used to calculate ground truth, and equation 6 was used for window prediction. The GRU-based model is likely to identify several frames besides the actual injected frame as weak anomalies because the injected frame might create several new (anomalous) CAN ID sequences.

$$X_g = \frac{Number\ of\ attack\ frames\ in\ T}{Total\ number\ of\ frames\ in\ T} \tag{3}$$

$$X_w = \frac{Number\ of\ weak\ anomalies\ in\ T}{Total\ number\ of\ frames\ in\ T} \tag{4}$$

$$Ground\ truth = \begin{cases} 1, X_g \geq \mathcal{E} \\ 0, X_g < \mathcal{E} \end{cases} \quad (5)$$

$$Window\ prediction = \begin{cases} 1, X_w \geq \mathcal{E} \\ 0, X_w < \mathcal{E} \end{cases} \quad (6)$$

5. EXPERIMENT RESULTS AND PERFORMANCE EVALUATION

A. Experimental Setting

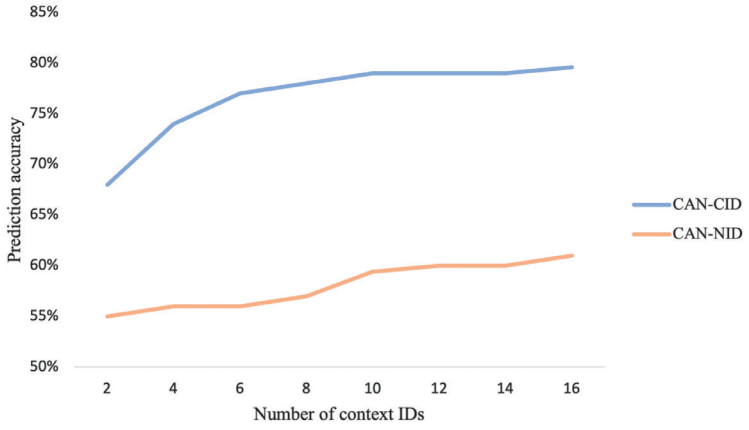
We selected ten benign datasets for training and two benign datasets for ID-based threshold estimation. To create the sliding window, ten IDs were selected from both sides of the target ID (n = 10). In addition, the attack datasets were split into 100-millisecond windows to identify attack windows, which can be considered smaller windows for near-real-time detection. This resulted in about 250 IDs per prediction window. To make the model more lightweight, only 32 GRU nodes were used in the hidden layer, followed by a 0.2 dropout layer. The ROAD, HCRL CH and HCRL SA datasets include 106, 27 and 45 nodes respectively (N). Based on a grid search, we observed that small anomaly thresholds (e.g. 0.01) work well with a large N and large anomaly thresholds (e.g. 0.1) work well with a small N. Hence, for the ROAD dataset, we set the anomaly threshold to 0.01, and for the HCRL datasets, the threshold was set to 0.1. A grid search was used for hyperparameter optimization, and the same parameters used in the ROAD dataset were also used for both HCRL datasets. We selected the best smallest hyperparameters for n, the number of GRU nodes and the embedding size. The proposed algorithm was implemented using Python 3.8 with TensorFlow and the Keras library. Experiments were run on a MacBook Pro 2.2 GHz Intel Core i7 with 16 GB RAM.

We compared CAN-CID with two baseline methods, that is, the N-gram-based model (N-gram) [18] and the transition matrix-based model (transition matrix) [19], where both models detect anomalies based on observed benign ID sequences. In addition, we used a variant of CAN-CID, referred to as CAN-NID (CAN Next ID prediction). CAN-NID is similar to CAN-CID, except the GRU model takes context IDs from one side (previous IDs). Optimized hyperparameters for the CAN-NID model include 16 previous IDs as the context, two hidden GRU layers with 128 nodes and a dense output layer with a softmax activation function. We also fine-tuned both baseline models for each dataset for a fair comparison with our model. To evaluate the model performance, we used F1-Score, false-positive rate (FPR) and false-negative rate (FNR) [33].

B. Results and Discussion

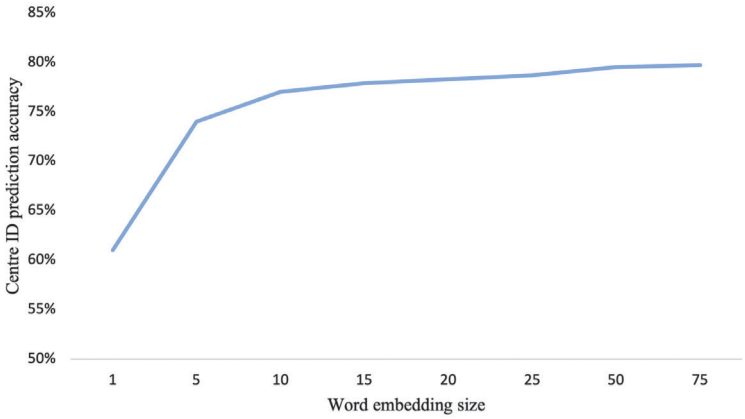
The detection accuracy of the GRU model of CAN-CID depends on the centre word prediction accuracy. We expect accurate predictions for benign frames and inaccurate predictions for attack frames to detect weak anomalies. To identify the optimum context from both sides of the centre ID, we tested various IDs by targeting the highest prediction accuracy for a sample from the benign dataset. Similarly, we tested a number of previous IDs for the CAN-NID GRU model. As shown in Figure 8, the CAN-CID model achieved 80% accuracy, whereas CAN-NID achieved a maximum of 61% accuracy for 16 context IDs. This highlights the effectiveness of the CBOW approach for CAN sequences. However, achieving 100% prediction accuracy is not realistic due to randomness incurred from jitters [28]. Considering the computational efficiency, we selected ten context words (79%) from each side for CAN-CID and 12 context words (60%) for CAN-NID.

FIGURE 8: COMPARISON OF CENTRE ID (CAN-CID MODEL) AND NEXT ID (CAN-NID MODEL) PREDICTION ACCURACY



Word embedding size is another critical factor for accuracy and computational efficiency. Therefore, we tested the CAN-CID model with different embedding sizes, as shown in Figure 9. We observed that accuracy improved up to an embedding size of 50. Therefore, we used 50 as the embedding size for both GRU models.

FIGURE 9: ACCURACY IMPROVEMENT WITH WORD EMBEDDING SIZE FOR THE CAN-CID MODEL



The F1-Scores, FPRs and FNRs of the CAN-CID and CAN-NID models and two baseline models are presented in Tables II and III for the ROAD dataset. Tables II and III report fabrication and masquerade attacks respectively, where the best performance (F1-Score) for each attack is shown in bold. As shown in the tables, the CAN-CID model outperforms the two baseline models for every attack and achieved a 100% F1-Score for six attacks. More importantly, this model achieved 0% or very small FPR and FNR values, which are critical aspects for an IDS. The CAN-NID model also outperformed baseline models for seven attacks. A fuzzing attack is relatively easy to detect due to illegal ID injection, and therefore, all models except the transition model achieved an F1-Score of 100%. However, correlated signal and correlated signal masquerade attack detection rates are low compared to other attacks. This might be because they target the second most frequent ID, which has a slightly random transmission rate compared to other IDs. Therefore, it creates more sequences, which results in more valid sequences being created, even for attack frames. This is a limitation of the proposed model, whereby it achieves a lower detection rate for attacks that target IDs with random transmission rates. However, a greater number of CAN IDs have fixed transmission rates [6], and therefore, CAN-CID can detect the majority of injection attacks. Further, since CAN IDs have fixed transmission rates, most ID sequences are likely to be independent of driving behaviours. This makes the model resilient to such changes. However, one of the limitations of the proposed model is that the CAN-CID model requires greater variety in the benign data to minimize the unseen CAN ID sequences and time intervals.

TABLE II: COMPARISON OF CAN-CID AND CAN-NID MODELS AND BASELINE MODELS DETECTION PERFORMANCE OF FABRICATION ATTACKS (ROAD DATASET)

Attack	Model	F1-Score	FPR	FNR
Fuzzing	Transition matrix N-gram CAN-NID CAN-CID	71% **100%** **100%** **100%**	48% 0% 0% 0%	0% 0% 0% 0%
Correlated signal	Transition matrix N-gram CAN-NID CAN-CID	90% 27% 78% **91%**	10% 0% 21% 2%	6% 100% 42% 12%
Max speedometer	Transition matrix N-gram CAN-NID CAN-CID	79% 89% **100%** **100%**	27% 0% 0% 0%	0% 28% 0% 0%
Reverse light on	Transition matrix N-gram CAN-NID CAN-CID	63% 87% 94% **100%**	57% 0% 1% 0%	0% 29% 2% 0%
Reverse light off	Transition matrix N-gram CAN-NID CAN-CID	92% 94% **100%** **100%**	9% 0% 0% 0%	0% 16% 7% 0%

TABLE III: COMPARISON OF CAN-CID AND CAN-NID MODELS AND BASELINE MODELS DETECTION PERFORMANCE OF MASQUERADE ATTACKS (ROAD DATASET)

Attack	Model	F1-Score	FPR	FNR
Correlated signal masquerade	Transition matrix N-gram CAN-NID CAN-CID	38% 27% 64% **89%**	10% 0% 22% 4%	86% 100% 57% 10%
Max speedometer masquerade	Transition matrix N-gram CAN-NID CAN-CID	79% 99% 86% **100%**	27% 0% 0% 0%	0% 1% 36% 0%
Reverse light on masquerade	Transition matrix N-gram CAN-NID CAN-CID	63% 87% 94% **99%**	57% 0% 1% 0%	0% 29% 2% 1%
Reverse light off masquerade	Transition matrix N-gram CAN-NID CAN-CID	92% 94% 95% **100%**	9% 0% 0% 0%	0% 16% 7% 0%

Figures 10 and 11 present comparisons of the attack detection performance of time-based and GRU models. Figure 10 shows fabrication attacks, whereas Figure 11 shows masquerade attacks. Typically, the time-based model is capable of detecting fabrication attacks with a higher F1-Score, whereas it fails to detect masquerade attacks. In contrast, the GRU model is capable of detecting both types of attacks with a higher F1-Score.

FIGURE 10: TIME-BASED AND GRU MODEL DETECTION PERFORMANCE FOR FABRICATION ATTACKS

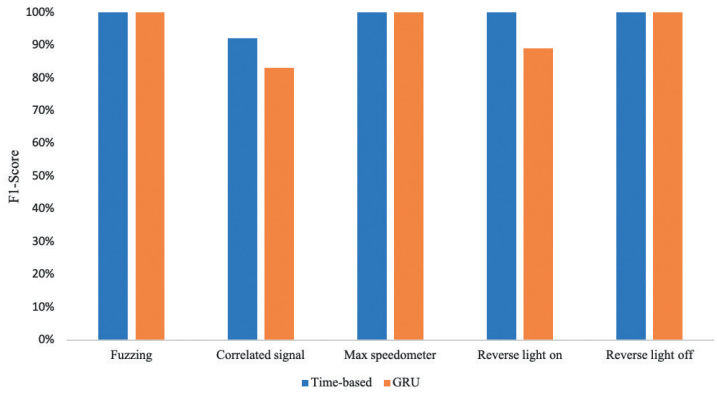
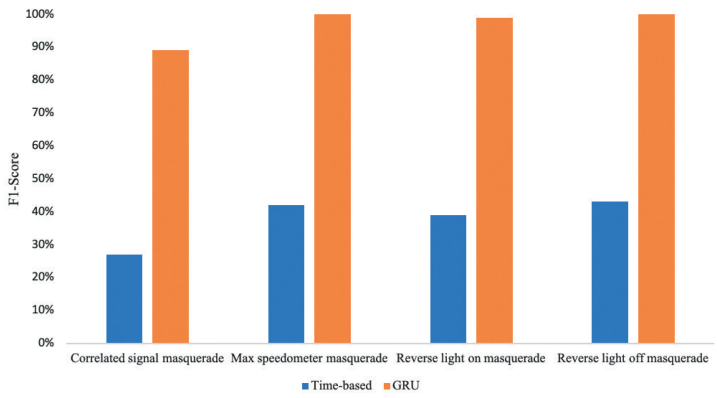


FIGURE 11: TIME-BASED AND GRU MODEL DETECTION PERFORMANCE FOR MASQUERADE ATTACKS



As mentioned earlier, we used two HCRL datasets to evaluate the generalization capability of the proposed model. The results from these two datasets are similar to the results we achieved from the ROAD dataset (see Table IV and V). The CAN-CID

and CAN-NID models outperformed both baseline models. However, both baseline models showed comparatively better results for the HCRL datasets. This might be due to the HCRL datasets having a limited number of IDs, thus limiting the number of CAN ID sequences created compared to the ROAD dataset, which would help achieve higher predictability.

TABLE IV: COMPARISON OF ATTACK DETECTION PERFORMANCE OF THE CAN-CID AND CAN-NID MODELS AND THE BASELINE MODELS FOR THE HCRL CH DATASET

Attack	Model	F1-Score	FPR	FNR
DoS	Transition matrix	75%	52%	0%
	N-gram	96%	10%	0%
	CAN-NID	97%	6%	0%
	CAN-CID	**99%**	1%	0%
Fuzzy	Transition matrix	91%	20%	0%
	N-gram	94%	14%	0%
	CAN-NID	97%	6%	0%
	CAN-CID	**100%**	0%	0%
Gear Spoofing	Transition matrix	98%	4%	0%
	N-gram	98%	4%	0%
	CAN-NID	99%	1%	1%
	CAN-CID	**100%**	0%	0%
RPM Spoofing	Transition matrix	86%	28%	0%
	N-gram	98%	4%	0%
	CAN-NID	**99%**	0%	2%
	CAN-CID	**99%**	0%	2%

TABLE V: COMPARISON OF ATTACK DETECTION PERFORMANCE OF THE CAN-CID AND CAN-NID MODELS AND THE BASELINE MODELS FOR THE HCRL SA DATASET

Attack	Model	F1-Score	FPR	FNR
Flooding	Transition matrix	89%	28%	0%
	N-gram	99%	2%	0%
	CAN-NID	**100%**	0%	0%
	CAN-CID	**100%**	0%	0%
Fuzzy	Transition matrix	85%	28%	0%
	N-gram	99%	1%	0%
	CAN-NID	99%	1%	0%
	CAN-CID	**100%**	0%	0%
Malfunction	Transition matrix	68%	54%	0%
	N-gram	84%	28%	0%
	CAN-NID	91%	2%	17%
	CAN-CID	**96%**	0%	4%

Detection latency is another criterion that we focused on improving as it is vital for moving vehicles. Table VI compares average detection latency for CAN-CID, CAN-NID and the two baseline models. The IDS monitors CAN traffic for 100 ms and gives the prediction in 10 ms. CAN-CID outperformed CAN-NID and the two baseline models. The small amount of time required for monitoring and prediction allows the vehicle driver or the vehicle itself to take appropriate countermeasures. Therefore, considering the detection capability and latency, the proposed algorithm is a practically deployable solution to detect cyberattacks on the CAN bus. Furthermore, using the CAN ID and time as the only features of the ensemble model improves the detection latency in a resource-constrained environment. Additionally, this model is likely to have a better generalization capability than a payload-based model as data (payload) specifications might change significantly across different vehicle makes and models.

TABLE VI: AVERAGE DETECTION LATENCY COMPARISON FOR A 100 MS PREDICTION WINDOW

Model	Detection latency (ms)
Transition matrix	36
N-gram	452
CAN-NID	12
CAN-CID	**10**

6. CONCLUSION AND FUTURE WORKS

Increased connectivity and complexity in modern automobiles create more attack surfaces that could allow attackers to take control of automobiles. Cyberattacks on moving vehicles are highly dangerous and could result in serious injury or even deadly consequences. Therefore, there is a dire need to implement defence mechanisms against these attacks. Due to the complexity of CAN data and the different characteristics of different types of potential attacks, this work demonstrates that the solution requires an ensemble model with an optimized model for each field of CAN data.

Hence, we proposed CAN-CID, a novel context-aware ensemble IDS for CAN bus security. Our experiments showed that the ensemble model improved the overall attack detection performance and outperformed two baselines and a variant of the proposed model. Additionally, the proposed CAN-CID model has a low detection latency,

which is necessary for a deployable in-vehicle IDS. We also identified potential future work to improve the model. We propose adding another model to the ensemble model to monitor the CAN payload and thus detect more advanced attacks, which would not change ID sequences or frequencies. Secondly, the IDS should be capable of adapting to new data. Therefore, we plan to work on introducing streaming learning capability. Finally, we plan to deploy the IDS and test it under real-world conditions. These additions to the proposed model will help keep moving vehicles secure from an even wider range of in-vehicle network cyberattacks.

REFERENCES

[1] O. Y. Al-Jarrah, C. Maple, M. Dianati, D. Oxtoby and A. Mouzakitis, 'Intrusion detection systems for intra-vehicle networks', *IEEE Access*, vol. 7, pp. 21266–21289, 2019.

[2] R. N. Charette, 'This Car Runs on Code'. Accessed: Nov. 28, 2021. [Online]. Available: https://spectrum.ieee.org/transportation/systems/this-car-runs-on-code

[3] D. Moller and R. Hass, *Guide to Automotive Connectivity and Cybersecurity: Trends, Technologies, Innovations and Applications*, Springer, Cham, 2019.

[4] O. Avatefipour and H. Malik, 'State-of-the-art survey on in-vehicle network communication (CAN-Bus) security and vulnerabilities', *IJCSN*, vol. 6, no. 6, pp. 720–727, 2017.

[5] Z. Cai, A. Wang, W. Zhang, M. Gruffke and H. Schweppe, '0-days & Mitigations: Roadways to Exploit and Secure Connected BMW Cars', in *Black Hat USA*, 2019.

[6] C. Miller and C. Valasek, 'CAN Message Injection', 28 June 2016. Accessed: Nov. 29, 2021. [Online]. Available: http://illmatics.com/can%20message%20injection.pdf

[7] S. Nie, L. Liu and Y. Du, 'Free-fall: Hacking Tesla from wireless to CAN bus', in *Black Hat USA*, 2017.

[8] M. Engstler, 'Heavy On Connectivity, Light On Security: The Challenges Of Vehicle Manufacturers', Jan. 15, 2021. Accessed: Nov. 29, 2021. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2021/01/15/heavy-on-connectivity-light-on-security-the-challenges-of-vehicle-manufacturers/?sh=68dda9247fc7

[9] F. Lambert, 'Tesla is challenging hackers to crack its car, and it is putting ~$1 million on the line', Jan. 10, 2020. Accessed: Nov. 29, 2021. [Online]. Available: https://electrek.co/2020/01/10/tesla-hacking-challenge/

[10] M. E. Verma, M. D. Iannacone, R. A. Bridges, S. C. Hollifield, B. Kay and F. L. Combs, 'ROAD: The Real ORNL Automotive Dynamometer Controller Area Network Intrusion Detection Dataset (with a comprehensive CAN IDS dataset survey & guide)', *arXiv preprint arXiv:2012*, vol. 14600, 2020.

[11] N. Salman and M. Bresch, 'Design and implementation of an intrusion detection system (IDS) for in-vehicle networks', M.S. thesis, Dept. Comp. Sci. Eng., Univ. Gothenburg, Sweden, 2017.

[12] Y. Xun, Y. Zhao and J. Liu, 'VehicleEIDS: A Novel External Intrusion Detection System Based on Vehicle Voltage Signals', *IEEE Internet of Things Journal*, 2021.

[13] A. Tomlinson, J. Bryans and S. A. Shaikh, 'Using a one-class compound classifier to detect in-vehicle network attacks', in *Proc. Genet. Evol. Comput. Conf. Companion*, 2018.

[14] S. Longari, M. Zago and S. Zanero, 'CANnolo: An Anomaly Detection System Based on LSTM Autoencoders for Controller Area Network', *IEEE Trans. Netw. Serv. Manag*., vol. 18, no. 2, pp. 1913–1924, 2021.

[15] P. Balaji and M. Ghaderi, 'NeuroCAN: Contextual Anomaly Detection in Controller Area Networks', in *IEEE Int. Smart Cities Conf. (ISC2)*, 2021.

[16] M. J. Kang and J. W. Kang, 'Intrusion detection system using deep neural network for in-vehicle network security', *PloS one*, vol. 11, no. 6, 2016, Art. no. e0155781.

[17] J. Zhang, F. Li, H. Zhang, R. Li and Y. Li, 'Intrusion detection system using deep learning for in-vehicle security', *Ad Hoc Netw*., vol. 95, 2019, Art. no. 101974.

[18] H. K. Kalutarage, O. M. Al-Kadri, M. Cheah and G. Madzudzo, 'Context-aware anomaly detector for monitoring cyber attacks on automotive CAN bus', in *Proc. – CSCS 2019: ACM Comp. Sci. Cars Symp*., 2019.

[19] M. Marchetti and D. Stabili, 'Anomaly detection of CAN bus messages through analysis of ID sequences', in *IEEE Intell. Veh. Symp. Proc. (IVS)*, 2017.

[20] D. H. Blevins, P. Moriano, R. A. Bridges, M. E. Verma, M. D. Iannacone and S. C. Hollifiel, 'Time-Based CAN Intrusion Detection Benchmark', *arXiv preprint arXiv:2101*, vol. 05781, 2021.
[21] A. K. Desta, S. Ohira, I. Arai and K. Fujikawa, 'ID Sequence Analysis for Intrusion Detection in the CAN bus using Long Short Term Memory Networks', in *2020 IEEE Int. Conf. Pervasive Comput. Commun. Workshops, PerCom Workshops 2020*.
[22] S. Tariq, S. Lee, K. H. Kim and S. S. Woo, 'CAN-ADF: The controller area network attack detection framework', *Comput. Secur*., vol. 94, 2020, Art. no. 101857.
[23] E. Aliwa, O. Rana, C. Perera and P. Burnap, 'Cyberattacks and Countermeasures for In-Vehicle Networks', *ACM Comput. Surv. (CSUR)*, vol. 54, no. 1, pp. 1–37, 2021.
[24] K. Koscher, A. Czeskis, F. Roesner, S. Patel, T. Kohno, S. Checkoway, D. McCoy, B. Kantor, D. Anderson and H. Shacham, 'Experimental security analysis of a modern automobile', in *Proc. 31st IEEE Symp. Secur. and Priv*., pp. 447–462, 2010.
[25] V. Chockalingam, I. Larson, D. Lin and S. Nofzinger, 'Detecting Attacks on the CAN Protocol With Machine Learning', in *Annu. EECS*, 2016.
[26] J. Dürrwang, J. Braun, M. Rumez, R. Kriesten and A. Pretschner, 'Enhancement of automotive penetration testing with threat analyses results', *SAE Int. J. Transp. Cybersecur. Priv*., pp. 91–112, 2018.
[27] S. Woo, H. J. Jo and D. H. Lee, 'A practical wireless attack on the connected car and security protocol for in-vehicle CAN', *IEEE Trans. Intell. Transp. Syst*., vol. 16, no. 2, pp. 993–1006, 2014.
[28] K.-T. Cho and K. G. Shin, 'Error handling of in-vehicle networks makes them vulnerable', in *Proc. ACM Conf. Comput. Comm. Secur*., 2016.
[29] H. M. Song, J. Woo and H. K. Kim, 'In-vehicle network intrusion detection using deep convolutional neural network', *Veh. Commun*., vol. 21, 2020, Art. no. 100198.
[30] M. L. Han, B. I. Kwak and H. K. Kim, 'Anomaly intrusion detection method for vehicular networks based on survival analysis', *Veh. Commun*., vol. 14, pp. 52–63, 2018.
[31] T. Mikolov, K. Chen, G. Corrado and J. Dean, 'Efficient estimation of word representations in vector space', in *1st Int. Conf. Learn. Represent., ICLR 2013 – Workshop Track Proc*., 2013.
[32] S. Yang, X. Yu and Y. Zhou, 'LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example', in *Proc. – 2020 Int. Workshop Electron. Commun. Art. Intell., IWECAI 2020*.
[33] O. Minawi, J. Whelan, A. Almehmadi and K. El-Khatib, 'Machine Learning-Based Intrusion Detection System for Controller Area Networks', in *DIVANet 2020 – Proc. 10th ACM Symp. Des. Anal. Intell. Veh. Netw. Appl*., 2020.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

Towards a Digital Twin of a Complex Maritime Site for Multi-Objective Optimization

Joseph A. J. Ross
University of Plymouth
joseph.ross@plymouth.ac.uk

Kimberly Tam
University of Plymouth

David J. Walker
University of Plymouth

Kevin D. Jones
University of Plymouth

Abstract: Her Majesty's Naval Base (HMNB) Devonport is a complex maritime site in Plymouth, United Kingdom (UK). Using digital twin technology, the authors will model and simulate the physical entity of the dockyard to optimize for a set of critical priorities. Digital twins are virtual representations of a physical entity, such as a vehicle. They can fully model a complex environment, accurately modelling individual layers within the entity, with each layer accessing data required from other layers. This results in an accurate simulation so that when changes are made in one layer of the model, the impact across the other layers may be observed. An end-user could interact with this digital twin to understand how changing input parameters would affect the measured outputs, allowing the end-users to simulate different options and compare the simulated outcomes before deciding a course of action. If the digital twin is of higher fidelity, the simulated outcomes would be more accurate and demonstrate potentially unintended effects allowing for a more comprehensive overview for the decision-maker. From this digital twin, a decision-maker can manually identify the best parameters to simulate the outcomes through the digital twin. However, using multi-objective optimization can reduce this process so that the twin can create the inputs, monitor the outcomes, and repeatedly try to produce a specific number of outcomes to choose from. These outcomes would be based on a few priorities initially set, and the optimizer would change inputs to enhance each of these priorities. At HMNB Devonport, three main priorities have been identified: cost reduction, time efficiency and carbon neutrality.

Keywords: *digital twin, visualization, multi-objective optimization, dockyard logistics*

1. INTRODUCTION

HMNB Devonport is a complex maritime site with many individual stakeholders. The site's purpose is to support the Royal Navy (RN) and its operations; to this end, the site requires more than purely service personnel. The Ministry of Defence (MOD) also assigns many civil servants to the site to manage and run various services across the site. However, public-sector employees are not sufficient to run and maintain the whole site, so a variety of contractors are required onsite, some of which are permanently based on site. Others are temporarily contracted to complete short-term projects. With all these separate organizations and smaller teams, there are many 'team leaders' and 'project managers', each focused on their priorities or role within the organization. The primary stakeholder is the MOD, and they own the freehold for much of the site. Another key stakeholder is Babcock International Group PLC (Babcock); Babcock owns the remainder of the site's freehold and holds the contract to maintain many of the RN's maritime assets. In each naval base, there is one stakeholder from within the RN, who is principally in charge on behalf of the MOD, the Naval Base Commander (NBC). Through conversation with HMNB Devonport's NBC, three key priorities have been identified related to all onsite operations. These are to minimize cost, improve efficiency and reduce environmental impact, aiming for net-zero emissions by 2050. Various government publications also mirror this. For example, Maritime 2050 [1] states that the UK will aim to be a role model in minimizing carbon emissions within shipping. The Department for Business [2] then states that emissions will be cut by 78% by 2035, adding that the long-term target has been enacted into legislation.

In addition to these priorities, Maritime 2050 [1] looks at technological progress as being vital for the future of the maritime industry, focusing on digitization and stating that 'the UK maritime sector will be "digital by default"' [1]. The MOD [3] has also given insight into the future of the RN, stating that investment will be focused on 'delivering a more modern, high-tech and automated Navy'. This process of digitization could begin by utilizing digital twin technology and optimization. A digital twin is an accurate digital model of a physical entity; in this case, it would be the dockyard, HMNB Devonport, and used to simulate operations to aid in risk management, planning maintenance and optimization. A digital twin would allow a complete understanding of the site's current performance and allow for a platform for planning and adapting future planning based on accurate simulated results. Optimization is the process of selecting the best solution based on a pre-determined criterion. There are many types of optimization; however, most have a single objective to optimize. For example, an objective may be to reduce the time it takes to do something, so the most optimal solution is the one that takes the shortest amount of time. One method of finding this solution is through evolutionary algorithms, which use natural evolution

as a concept to improve solutions. This starts with a set of random solutions, which are then simulated and their effectiveness measured and given a fitness score. The highest scored solutions will then be used to create a new set, which will be simulated and rated, and subsequently compared with the original set, and the lowest scorers will be removed. This process will iterate many times, with mutations occurring where the solutions can continue to develop some new information to be simulated. This will finally result in a single solution that has the highest rating. If the objective is to reduce time, this would be the solution that results in the shortest time required. However, in many real-life problems, there is more than a single objective. It may be the case that the shortest solution would, in practice, cost much more than the second shortest, in which case the decision-maker may choose to compromise between time and cost.

This would be a multi-objective problem that would require three objectives to be considered, and multi-objective optimization algorithms would provide multiple solutions showing different compromises between the objectives. This digital twin could then work hand in hand with multi-objective optimization algorithms to have the digital twin provide the inputs for the model and then adapt those inputs over time until the best few solutions are found for the pre-defined priorities. The user would then be able to make an informed decision based on the results from many simulations, with each new set of inputs being created based upon the learning from the previous inputs. This would then provide solutions that take all priorities into account, while still allowing the user to select a solution according to their own focus. The use of digital twins in this way would not be limited to HMNB Devonport. It would also apply to other complex maritime sites, both military and commercial, and would be a basis for creating smart ports, where the digital twins could be formed using live data from around the port.

2. DIGITAL TWIN

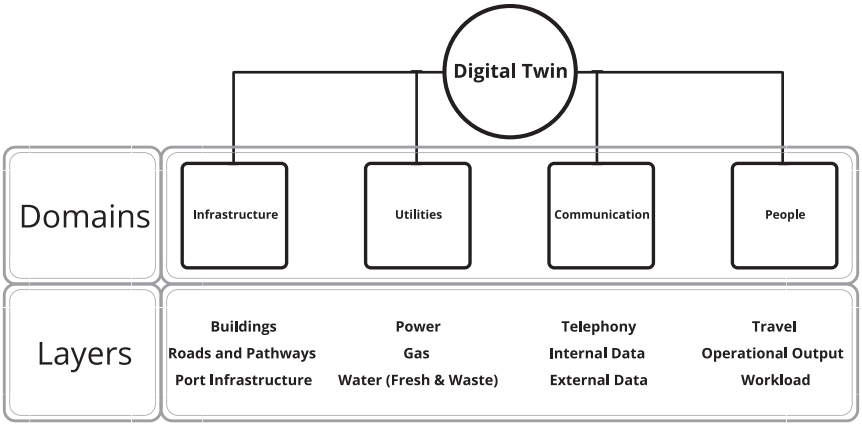
Modern digital twin technology is being used in a wide variety of industries and for a number of different purposes. Tao *et al*. [4] and VanDerHorn and Mahadevan [5] both observed and commented on this rise of popularity, stating that digital twins were gaining interest from both academia and industry, evidenced by a notable increase in publications on and patents for digital twin technologies. The most common applications of digital twins so far have been in the manufacturing industry [6], [7] and the design and monitoring of smart cities [8], [9]. However, this concept is not limited to these two areas and can be applied in a plethora of challenges across all industries. Glaessgen and Stargel [10] discuss applications within the automotive industry, specifically for NASA and the US Air Force, to combine simulation with the vehicle's onboard computer to predict the life of the vehicles and provide additional

levels of safety and reliability. Pylianidis *et al*. [11] introduce how digital twins can be applied to agriculture and Braun [12] demonstrates how digital twins can be used in medicine. Damiani *et al*. [13] show that digital twin technology has already been used in a complex maritime terminal, specifically the Port of Genoa; however, the digital twin they are discussing models the port for the purpose of supporting energy management.

Digital twin technology provides a virtual environment that simulates a real-world entity. A single digital twin contains multiple domains and layers so that when viewed holistically, the virtual entity accurately reacts just as the physical entity would. This simulation, therefore, allows the users to modify and observe data across virtualized domains and then visualize the effects. It is essential to realize that any dockyard digital twin user in this study may vary from highly technical contractors, civil servants, or service personnel looking to identify and plan the future of the site to non-technical assistants who could be MOD apprentices or junior service personnel with no technical knowledge of the site, computing, or engineering. These challenges mean that it is vital for any proposed digital twin to be visually understandable for any potential user while also providing enough valuable information to aid technical users in their planning. This suggests that any solution will need to understand both the context or environment and the users' needs to provide both the right level or type of detail and the right amount of those details.

To model HMNB Devonport using a digital twin, the authors propose using four separate domains, each with three layers. This is shown in Figure 1.

FIGURE 1: DOMAINS AND LAYERS OF THE DIGITAL TWIN



A. Infrastructure

HMNB Devonport is a large maritime site with over 650 buildings covering more than 650 acres of land [14]. The buildings layer would replicate the physical layout of buildings within the site and would then present an overlay for other layers to give additional information. Buildings will be replicated to scale, accurately showing size and shape to represent the site visually. Additionally, the buildings layer would include data regarding the use of the building for application within other layers. This layer would be the basis of the user interface, as the user would visualize the other layers around a well-known map of the site, which would immediately make the software more intuitive.

Another layer based on the physical infrastructure of the site is the roads and pathways. This layer would be vital for monitoring movements on the base and highlighting additional problems associated with a road being closed due to maintenance or an unplanned closure due to an act of God. This layer would not only show where the roads and pathways are but also the direction of travel and capacity for certain vehicles. This would then allow routing applications to run to monitor efficiency. This layer would identify single points of failure, showing routes that would not be able to occur if one road or path became unavailable or blocked. This would inform future developments to avoid single points of failure as a simple blockage could have a time-consuming and expensive effect on the site, which would be detrimental to two of the base's three priorities.

The site also has operational infrastructure, such as the dockyard basins, wharves, and docks, as well as locations where operational equipment (e.g., cranes, forklifts, and other dockyard equipment) is stored. By having these locations as a layer in the digital twin, routing can take place to see where physical assets must come from and go. This would facilitate the measurement of carbon emissions for different transfer scenarios as well as storage optimization across dockyard infrastructure, showing where equipment should be stored or where the work should take place. This could also go a step further to identify where ships and boats should be moored based on the work required to make the process more efficient. Because of the level of detail across these layers, it is then possible to approach a challenge with multiple objectives and optimize to any set of any number of objectives (see Section 3). Meeting multiple objectives through planning may involve dockyard plant and assets being transported less. This would minimize emissions, onsite traffic and congestion and save time, as equipment would not need to be moved around the site before commencing work. Workers would not need to be paid to move the equipment, and the equipment would not be unavailable for use elsewhere. This reduction in time would also reduce the overall financial impact of the work, as workers would be required to spend less time on the project.

B. Utilities

Power is required all over HMNB Devonport for almost all operations. As shown in Figure 1, the communications layers are entirely reliant on power. Without power, all communications layers would cease to function. The infrastructure layer would also be somewhat affected – roads and pathways would still be available; however, signalling and access (e.g., gates, electronic door access) may be negatively affected. This could result in people being locked in or out of buildings or areas if they are designed to fail closed or could present a security risk if they are designed to fail open. Affecting offices and workplaces, therefore, also influences the people layer. It is vital that planned and unplanned power outages are monitored so that the site can predict what would be affected. The digital twin would do this inherently – with a power layer, when an outage is simulated, the twin would be able to identify where systems are affected. For example, a power outage from the utility domain would affect certain buildings from the infrastructure domain. This could then prevent a network switch from the communications domain from functioning, which in turn could mean that the internal data layer (communications domain) would become unavailable for a portion of the site. That information would be passed to the buildings layer (infrastructure domain), showing where the network would not be available and then identifying which operations cannot take place; alternatively it would allow for planning to mitigate the risk or plan to move operations around so that the highest priority operations can take place.

Similar to mains power around the site, the gas mains would be visible on the gas utility layer for the digital twin, and this would indicate which buildings are connected to the gas mains. If there was a problem somewhere in the network, the digital twin would identify which areas would experience a disruption to the gas supply. This would highlight where vulnerabilities could be expected within the network and if an operation required gas, an alternative source could be planned.

Finally, for utilities in this particular study, water would be shown on the digital twin in its own layers, including clearwater, greywater and blackwater [15]. Clearwater is used all over the site for domestic purposes, such as drinking water or making tea or coffee in the restrooms and cleaning facilities. Clearwater must be safe for consumption and should be free from common bacteria such as legionella, which can be ensured by the correct storage of water and routine monitoring. In addition to this, clearwater is also vital for safety, as HMNB Devonport is a nuclear site and the decommissioned nuclear submarines require a constant supply of power and water to avoid overheating [16]. Considering the potentially severe consequences if power and water are not available at the dockyard, a dockyard like this is likely to benefit from a multiple-objective optimizer, as it will ensure, for example, lowering emissions without compromising safety.

These layers would also consider the source of the utility to be able to monitor effects outside of the dockyard. Any external water, gas or power supply would be integrated into the digital twin as well as onsite supplies, such as power generators.

C. Communication

The copper telephony network on site is vital for communications. Many people have migrated over to Voice over Internet Protocol (VoIP) telephony. However, many still require the copper network, and due to the age of this network, failures are not uncommon. This layer will show the distribution of copper cabling throughout the site to identify single points of failures that could cause outages over the site. This layer would also aid with planning for new cabling to be installed in the most effective way to improve efficient routing and to link with other infrastructure and works taking place.

In addition to the copper telephony network, the site has a variety of internal data networks which are air-gapped and provide local access to specific devices. Due to security, these networks must be completely isolated from each other and, in many cases, will go to the same buildings as different access permissions may be necessary for the same location. As a military site, security is of paramount significance, and any action must be compliant with the MOD's extensive security policies. The digital twin would also be able to compare existing networks and future planning within these security policies to ensure compliance and highlight any potential security vulnerabilities. Having these networks mapped onto separate layers would again make it possible to identify single points of failure. With the constant development of the site, it would also make it possible to plan future networks to be most efficient using the current infrastructure. This would allow for possible connections in any building, rather than the 'as required' request process the site currently has, as that consistently requires additional works to take place as operations move between buildings. When this is not centrally coordinated, this modular approach means that the site is regularly being worked on; however, the pit and duct network could be designed most optimally once and then installed in one action to avoid this problem.

The external network would also be a layer in the communications domain of the digital twin. If there was an issue externally that might reduce access to certain parts of our internal network, having the external connection mapped accurately would allow the site to predict problems based on external issues such as power outages.

D. People

The site's workforce is made up of thousands of personnel, with thousands of additional visitors each month onsite [17]. Without these people, the dockyard would not be operational. It is unclear whether other dockyards face similar circumstances,

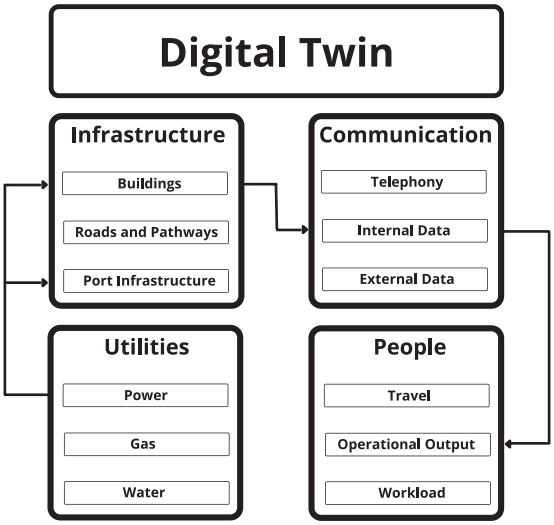
but only a portion of the employees at HMNB Devonport live in Plymouth, the city the dockyard resides in. The remainder commute in, so it is vital to consider how staff complete their commute. The travel layer will be associated with employees to show how road traffic and public transport delays or cancellations would affect the operational output of the site. A flexible digital twin would be able to model this accurately for this research and adapt to other dockyards or new situations if the pattern of workers changes over time. Today, many people may have other options for commuting, but some may not be able to attend work if there are unavoidable delays or cancellations. This aspect of the people layer will monitor the possible effects of worker subsets who may not be able to commute to the areas in which they work. Some employees may be able to work remotely; however, personal abodes will initially be outside the scope of this digital twin at this point.

The site could also be broken into several operational outputs. Teams, workers, and buildings would then be associated with these outputs so that people can be monitored in a way that allows the digital twin to identify what outputs may be reduced due to certain people not having access to the site or not having access to a utility or network in their building due to an outage. This would then be a factor that the digital twin model would produce to identify the severity of an outage, judging by how outputs would be affected.

E. Interconnected Layers and Domains

For the digital twin to be an effective tool, layers and domains will need to share data. Figure 2 shows a simulation of a power outage – the power layer in the utilities domain would show the outage, and that would be linked with the buildings layer within the infrastructure domain, so the buildings layer could identify which buildings have lost power due to the outage. From here, the other domains would access information from the buildings layer to monitor the effect on any other layers. For example, the internal data layer within the communications domain contains active networking equipment that requires power distributed throughout the site; if this power were to be removed, then some of the surrounding buildings from the infrastructure domain would lose power, and any active communications equipment would shut down, which would result in reduced connectivity for those working in those buildings, and also for those working elsewhere but requiring network traffic through that location. Additional layers would also be required outside of the domains to define the relationships between the domains and their layers. In Figure 2, this is represented by the external directed connectors between the domains. Processes within the digital twin would then allow for interlocking communication between domains, which would render a more accurate model of the characteristics of the dockyard.

FIGURE 2: SHARING DATA BETWEEN LAYERS AND DOMAINS



3. MULTI-OBJECTIVE OPTIMIZATION

Optimization algorithms allow users to solve problems in the most efficient way. An optimization problem includes a pre-defined model, which computes inputs to simulate a real-world problem and produce an output [18]. This outcome can then be compared with the previous outcomes, and the algorithm can try again. With a single objective, the algorithm will iterate for a significant number of inputs and then present one solution that creates the most favourable outcome.

Multi-objective problems are problems that require optimization for more than one objective; the technique above is not possible when there is more than a single objective. At HMNB Devonport, three main objectives have been identified: (1) minimize financial cost, (2) reduce time and maximize efficiency, and (3) reduce carbon emissions and improve sustainability. Additionally, pre-existing goals like physical safety and policy compliance are still around and would also need to be achieved within any possible solution.

A. HMNB Priorities

Financial cost is a crucial factor in the decision-making process for almost any operation, commercial and government alike. In general, budgets are provided for each project and compromises are needed to meet the budget. Multi-objective optimization

could only show solutions below a specific cost or present a variety of solutions that compromise on the three objectives but take all three into account. The efficiency of HMNB Devonport is vital for supporting the RN. The role of the dockyard is to ensure that a given number of ships are serviceable for war, should the order be given. If the time that a ship is unserviceable can be reduced, this would allow the ship to return to sea sooner, meaning that more ships would be 'ready' should the need arise. On projects where time is a factor, deadlines can be provided to the model, so only solutions that take a certain time are shown, or the best compromises could be shown, showing the most efficient solutions while taking cost and emissions into account. Sustainability is a big issue for governments and organizations around the world; the UK Government has declared that it intends to aim for a 78% reduction in carbon emissions by 2035 [1]. Plymouth City Council and ICE UK Ltd [19] report that HMNB Devonport is the largest energy consumer in the area and potentially in the whole of Plymouth. This shows that HMNB Devonport has a long way to go to reduce carbon emissions and that this needs to be a significantly weighted objective when it comes to any projects taking place. Environmental impact can be measured through the simulation, and solutions with the lowest impact that are also optimized for cost and time will be presented to the decision-maker.

With the three HMNB objectives presented above, one of the challenges is how divergent the solutions can be. In other words, completely optimizing one of these objectives (e.g., emissions) would likely be sub-optimal for the others (e.g., time and cost). Therefore, for the multi-objective problem presented here, all tasks are likely to have some compromises, which would create a set of solutions that would be balanced across each objective [20]. That said, this is highly dependent on the tasks and solutions available. Some decisions may require more compromise than others, and if people are aware of the compromises being made, they may create solutions that help satisfy multiple objectives more easily. For example, while the infrastructure for supply power may not change in the dockyard, the availability of renewable energy elsewhere could make decisions optimizing cost, emissions, and constant supply easier. All this information would then allow an informed decision based on their weighted priorities, while the solutions would have already taken the other priorities into account.

In order to combine digital twin technology and machine learning around multi-objective optimization, the following subsections discuss the most likely candidates for learning algorithms. These will be able to take a digital twin as input with its accurate representations and a set of weighted objectives defined by several different users. The output of this would then be a set of solutions to meet those objectives, complete with compromises, which the user can use, referring to the digital twin when needed, to make a well-informed decision.

B. Optimal Solutions

In Section 2, the authors showed how the complexity of a site, such as a dockyard, cannot operate effectively if only optimizing one goal. Moreover, contemporary issues (e.g., carbon emissions) on top of pre-existing challenges, such as fleet readiness, make decision-making even more challenging. This calls for machine learning and artificial intelligence to narrow all possible solutions to a smaller set of viable optimized solutions.

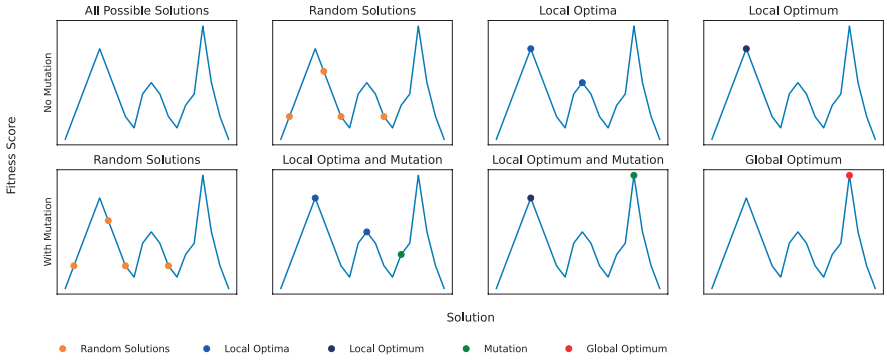
This set of balanced solutions across a number of objectives is called the Pareto optimal set, with each solution called a Pareto optimal solution [20]. Pareto optimal trade-off [21] can then occur where improvements can be made to optimize an objective further, and then at least one other objective must become less optimized. This would then allow the decision-maker to deviate from the solution deemed as the best compromise to meet the needs of the task for them, unlike other methods of multi-objective optimization. The Pareto approach treats the objectives separately; other methods use a weighted approach to prioritize the objectives into a single new objective, which can then be individually optimized for that weighting [22]. The Pareto approach, on the other hand, ensures that appropriate weighting is given to each objective to prevent decision-makers from routinely abandoning priorities for the site, which would prevent the site from, for example, meeting Government carbon emission targets [1], [2].

C. Evolutionary Algorithms for MOPs

As mentioned previously, a digital twin must evolve as the reality it simulates evolves. Evolutionary computing is based on the principles of natural evolution [18]. Evolutionary algorithms can be used to optimize multi-objective problems by starting with a random set of solutions. The chances are that these solutions would not contain the most optimized solution. The algorithm would run the solutions through the model to determine how optimized each solution would be. The algorithm would then create a new set by merging the best of the previous set to see if that has a positive effect. It is at this stage that mutations would also take place. Mutations throw new data into the set, which would allow the algorithm to explore more areas than defined within the initial random set of solutions. A local optimum would exist where moving in any direction would make the result worse. However, there may be others where the optimum is better than the original local optimum. When running an optimization algorithm, it would be possible for the initial set of solutions to not be near the actual global optimum – the most optimum of all the local optimums – if this were to happen without mutations taking place, it would not be possible to find the optimum. Figure 3 – created by the authors to demonstrate the purpose of mutation within evolutionary optimization algorithms – visualizes this. The initial set of solutions is shown by the squares, and over time they all localize on the perceived optimum, which

is the local optimum on the left. The global optimum on the right is only identified in the lower example, which included mutation, allowing the algorithm to spread out further after the initial random set is created to find the global optimum.

FIGURE 3: EVOLUTIONARY OPTIMIZATION WITH AND WITHOUT MUTATIONS: LOCAL AND GLOBAL OPTIMUMS



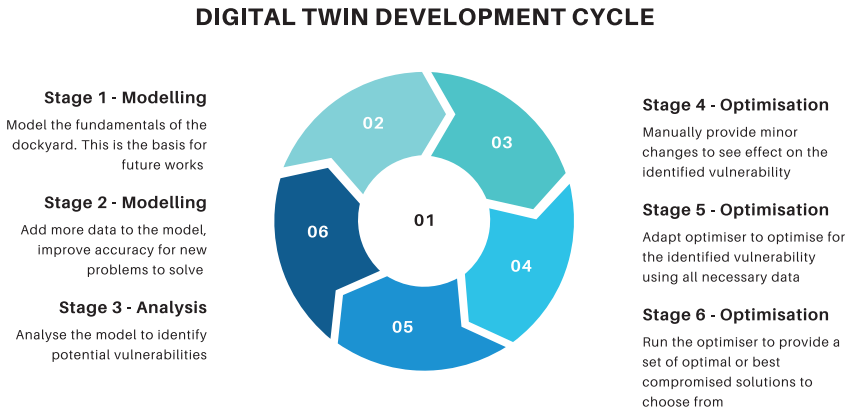
The airport gate assignment problem is a complex multi-objective optimization problem [23]; this problem is complex due to the ever-changing nature of an airport. This optimization requires dynamic optimization because live data is required to be able to produce the most optimal solutions, and as the data is frequently changing, the solutions will have to adapt. Changing solutions will also be part of the objectives, as the objectives are to maximize the airport's operational efficiency and convenience for passengers, which are two objectives that would not go hand in hand. This problem could be solved within a digital twin, which could be set up to take live data into account and then constantly run multi-objective optimization algorithms to find the Pareto optimal solutions. Through using a digital twin, a system would be created which would visualize the airport, planes and passengers while also being able to take routing and delays into account. Kaliszewski *et al*. [24] also discussed how multi-objective optimization using evolutionary computing could successfully optimize solutions using live data in the example of the airport gate assignment problem.

4. CONCLUSION AND FUTURE WORKS

A bespoke digital twin would allow the accurate simulation and modelling of HMNB Devonport as a complex site and would allow for in-depth visualization to facilitate access to the platform for any user, regardless of their technical expertise. The analysis of the digital twin would allow the user to identify potential vulnerabilities within the site and would also aid in planning future projects and minimizing vulnerabilities.

The next step for this research is to develop a prototype digital twin to act as a proof of concept for the dockyard. This will be in three distinct phases as shown in Figure 4 – (stages 1 and 2) modelling, (stage 3) analysis and (stages 4, 5 and 6) optimization. First, the digital twin must be designed to accurately model the site. Second, the digital twin will then analyse the model to identify any vulnerabilities, such as single or double points of failure. Finally, the digital twin will be adapted to use multi-objective optimization to find solutions for MOPs within the base, both those identified by the digital twin and externally inputted problems to be solved.

FIGURE 4: DIGITAL TWIN DEVELOPMENT CYCLE



This proof of concept will be compiled from gathered data, allowing for manual interaction with the data and for real-world users to keep the digital twin up-to-date. However, this could be improved for end-users by implementing a live data collection system utilizing Internet-of-Things (IoT) technology during the compilation of the digital twin model, which would both increase the accuracy of the model and efficiency, as end-users would not be required to manually interact with the raw data. Over time, this proof of concept will expand and tackle more and more problems, which will produce an estimation of the scale of the savings to be made – financial savings, environmental savings and increased efficiency – all while remaining compliant with policy and meeting the specific goals of the problem or potential vulnerability.

The level of detail in each layer could be increased so the digital twin could provide solutions to more MOPs. The number of MOPs that could be solved is almost limitless, so there will always be room for development in this regard. The digital twin could be expanded to model, analyse, and optimize other complex maritime sites, such as other naval bases (HMNB Faslane, HMNB Portsmouth) or even commercial ports. The technology also has potential uses for optimization in other industries, such as

hospitals or schools, which would use data and operations differently but would still use the analysis to identify vulnerabilities and optimization to reduce spending and improve efficiency.

5. ACKNOWLEDGEMENTS

We would like to acknowledge the Ministry of Defence for reading through the paper and checking for classification conflicts, and for funding the research through a doctoral studentship.

REFERENCES

[1] Department for Transport, "Maritime 2050: Navigating the future," 2019. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/872194/Maritime_2050_Report.pdf

[2] Department for Business, "UK enshrines new target in law to slash emissions by 78% by 2035," 2021. [Online]. Available: https://www.gov.uk/government/news/uk-enshrines-new-target-in-law-to-slash-emissions-by-78-by-2035

[3] Ministry of Defence, "Defence in a competitive age," 2021. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/974661/CP411_-Defence_Command_Plan.pdf

[4] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital Twin in Industry: State-of-the-Art," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2405–2415, 2019, doi: 10.1109/tii.2018.2873186.

[5] E. VanDerHorn and S. Mahadevan, "Digital Twin: Generalization, characterization and implementation," *Decision Support Systems*, vol. 145, p. 113524, 2021/06/01/ 2021, doi:10.1016/j.dss.2021.113524.

[6] Y. Lu, C. Liu, K. I. K. Wang, H. Huang, and X. Xu, "Digital Twin-driven smart manufacturing: Connotation, reference model, applications and research issues," *Robotics and Computer-Integrated Manufacturing*, vol. 61, p. 101837, 2020/02/01/ 2020, doi:10.1016/j.rcim.2019.101837.

[7] Q. Qi, F. Tao, Y. Zuo, and D. Zhao, "Digital Twin Service towards Smart Manufacturing," *Procedia CIRP*, vol. 72, pp. 237–242, 2018/01/01/ 2018, doi:10.1016/j.procir.2018.03.103.

[8] E. Shahat, C. T. Hyun, and C. Yeom, "City Digital Twin Potentials: A Review and Research Agenda," *Sustainability*, vol. 13, no. 6, p. 3386, 2021. [Online]. Available: https://www.mdpi.com/2071-1050/13/6/3386.

[9] J. E. Taylor, G. Bennett, and N. Mohammadi, "Engineering Smarter Cities with Smart City Digital Twins," *Journal of Management in Engineering*, vol. 37, no. 6, p. 02021001, 2021, doi:10.1061/(ASCE)ME.1943-5479.0000974.

[10] E. Glaessgen and D. Stargel, "The Digital Twin Paradigm for Future NASA and U.S. Air Force Vehicles," in *53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, 2012.

[11] C. Pylianidis, S. Osinga, and I. N. Athanasiadis, "Introducing digital twins to agriculture," *Computers and Electronics in Agriculture*, vol. 184, p. 105942, 2021/05/01/ 2021, doi:10.1016/j.compag.2020.105942.

[12] M. Braun, "Represent me: please! Towards an ethics of digital twins in medicine," *Journal of Medical Ethics*, vol. 47, no. 6, pp. 394–400, 2021, doi: 10.1136/medethics-2020-106134.

[13] L. Damiani, R. Revetria, and E. Morra, "A digital twin for supporting energy management in complex maritime terminals," presented at the International Maritime Transport and Logistics Conference, Alexandria, Egypt, 2019.

[14] Royal Navy. "HMNB Devonport." https://www.royalnavy.mod.uk/our-organisation/bases-and-stations/naval-base/devonport (accessed 2021-10-22).

[15] R. Brain, J. Lynch, and K. Kopp, "Defining Terms: Greywater, Blackwater and Clearwater," *USU Extension Publication*, 2015. [Online]. Available: https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1001&context=extension_curnat

[16] P. Webber, "Devonport and nuclear submarines: what are the risks?," *Scientists for Global Responsibility*, vol. 2017, no. 45, pp. 18–19, 2017. [Online]. Available: https://www.sgr.org.uk/sites/default/files/SGRNL45_Nuclearsubmarines.pdf.
[17] (2021). *One Devonport Infographic*.
[18] A. E. Eiben and J. E. Smith, *Introduction to evolutionary computing*, 2 ed. Berlin: Springer-Verlag, 2015.
[19] Plymouth City Council and ICE UK Ltd, "Business Feasibility Study for an Energy Services Company in Plymouth," 2010.
[20] X. Xu, Y. Tan, W. Zheng, and S. Li, "Memory-Enhanced Dynamic Multi-Objective Evolutionary Algorithm Based on Lp Decomposition," *Applied Sciences*, vol. 8, no. 9, p. 1673, 2018, doi: 10.3390/app8091673.
[21] B. Qu, "Evolutionary algorithms for solving multi-modal and multi-objective optimization problems," Doctoral, Nanyang Technological University, Singapore, 2011.
[22] A. A. Freitas, "A Critical Review of Multi-Objective Optimization in Data Mining: a position paper," *SIGKDD explorations*, vol. 6, no. 2, pp. 77–86, 2004.
[23] A. Bouras, M. A. Ghaleb, U. S. Suryahatmaja, and A. M. Salem, "The Airport Gate Assignment Problem: A Survey," *The Scientific World Journal*, vol. 2014, pp. 1-27, 2014, doi: 10.1155/2014/923859.
[24] I. Kaliszewski, J. Miroforidis, and J. Stańczak, "The airport gate assignment problem – multi-objective optimization versus evolutionary multi-objective optimization," *Computer Science*, vol. 18, no. 1, pp. 41–52, 2017, doi: 10.7494/csci.2017.18.1.41.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

The Design of Cyber-Physical Exercises (CPXs)

Siddhant Shrivastava
iTrust Centre for Research in Cybersecurity, Singapore University of Technology and Design

Francisco Furtado
iTrust Centre for Research in Cybersecurity, Singapore University of Technology and Design

Mark Goh
iTrust Centre for Research in Cybersecurity, Singapore University of Technology and Design

Aditya Mathur
iTrust Centre for Research in Cybersecurity, Singapore University of Technology and Design

Abstract: This paper explores the objectives, tactics, and strategies for identifying, planning, conducting, and evaluating an international cyber-physical exercise (CPX). The goal of a CPX is to improve defense capabilities for defending national critical infrastructure via global coordination. Lessons about CPX have been derived from a series of annual cyber-physical defense exercises conducted since 2015, referred to as Critical Infrastructure Security Showdowns (CISS). The cyber range of a CISS consists of a realistic and operational enterprise network coupled to water treatment and distribution plants in the form of physical testbeds and digital twins. These systems simulate and integrate information technology (IT) and operational technology (OT) scenarios that are ubiquitous in modern-day critical infrastructure controlled by industrial control systems (ICS). Participants from the red, blue, green, and white teams are assigned specific roles to attack, defend, visualize, and manage the plant, respectively. Each of these roles is evaluated via a specific set of metrics by a panel of judges and automated systems. The scoring criteria incentivize the red teams to design and launch novel attacks to contribute to and improve the cybersecurity community's knowledge base regarding offense and defense. The lessons distilled from these positive-sum games are analyzed and shared in the form of post-event reports.

From 2015 to 2021, CISS has constantly evolved to mimic contemporary cyber-physical security scenarios in the real world. These evolutions have forced the CISS organizing team to adapt and design novel infrastructure to support the changing needs of the event and its stakeholders, from tooling, logistics, and network infrastructure to scoring criteria and cross-disciplinary collaboration.

The cited reports on techniques, tactics, and procedures will be valuable to stakeholders from the military, industry, government, and academia.

Keywords: *cyber exercise, cyber-physical testbeds, remote organization, telepresence, digital twins, critical infrastructure security*

1. INTRODUCTION

Cyberattacks surged 151% in 2021 amid COVID-driven digitalization, with an average of 270 attacks per organization, costing US $3.6 million per company [1]. Furthermore, each company required 280 days on average to identify and respond to a cyberattack. The rise in the number of cyber-physical attacks, such as those against Colonial Pipeline, Florida Water Treatment Facility, and JBS Foods, highlighted significant security risks to our civilian critical infrastructure [2]. Thus, the need to "keep moving" in security preparedness has never been greater. Cyber defense exercises (CDXs) conducted worldwide have achieved significant milestones towards achieving international cooperation via deliberate practice in sharing knowledge about the tactics, techniques, and procedures for ensuring the cybersecurity of nations [3]. However, for several reasons, many such exercises had narrowed their scope to information technology (IT) networks and systems. Meanwhile, a new wave of cyber-physical attacks has expanded the threat of compromising operational technology (OT). But, taking note of this trend, organizations such as NATO's Cooperative Cyber Defense Centre of Excellence (CCDCOE) in its international Locked Shields and Crossed Swords exercises have adopted newer cyber-physical critical infrastructure as special systems (SS) [4]. For these types of exercises to become as popular and mainstream as regular cyber exercises, the defense community must communicate the lessons learned and share the anecdotal experiences of conducting cyber-physical exercises (CPXs). This paper serves as a CPX playbook [5] by laying the groundwork for identifying, planning, conducting, and evaluating CPXs for a variety of critical infrastructures.

The remainder of this paper is organized as follows. Section 2 delves into cyber-physical systems, cyber-physical attacks, and critical infrastructure and looks at the similarities and differences in their security in contrast to traditional cybersecurity. Section 3 discusses a specific CPX, based on water infrastructure, known as the Critical Infrastructure Security Showdown (CISS). Section 4 provides a playbook for conducting a CPX in different settings (physically on-site versus remote, using a hybrid approach with physical testbeds and digital twins, and multiple interconnected,

interdependent infrastructures, etc.). Section 5 provides a conclusion for the various stakeholders (the red, green, blue, white, and yellow teams) involved in such an event.

2. CYBER-PHYSICAL SYSTEMS SECURITY

A. Cyber-Physical Systems

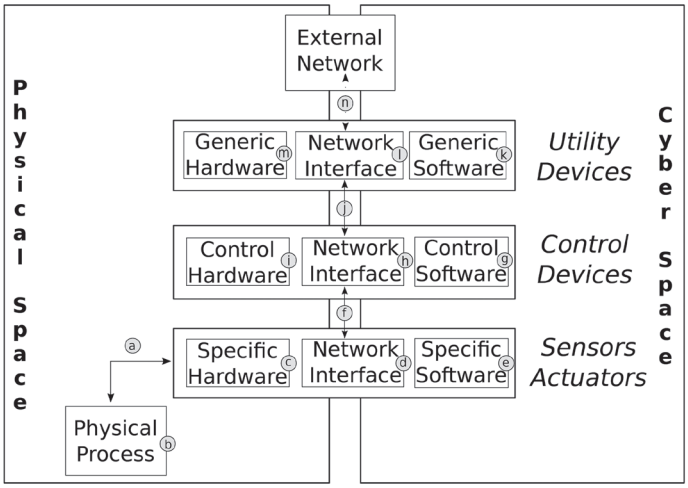
Cyber-physical systems (CPS) encompass all possible systems that exist in the cyber domain and the physical domain. As shown in Table I, examples include intelligent transportation systems, water treatment and distribution systems, automobiles, pacemakers, smartphones, and the power grid.

TABLE I: EXAMPLES OF CYBER-PHYSICAL SYSTEMS

Cyber-Physical System	Sensors / Actuators	Control devices	Utility devices
Water treatment SCADA system	Level sensor, Pumps, Flow sensor	PLCs	SCADA workstation, Historian server
Pacemaker	Pulse monitor, Pulse amplifier	Timing controller	Mobile application, Pulse logger
Automobile	Odometer / Throttle	ECUs using CAN Bus, i2c, FlexRay	Dashboard computer, mobile application

CPSs use sensors, actuators, and controllers connected via computing, networking, and physical processes to interact with entities and processes in the physical domain. The physical processes are integrated, monitored, and controlled by computers, known as controllers. The intelligence programmed in the cyber domain decides the steps that the physical processes should take, given the state of the system. This enables automation, control, and quality assurance of processes, which would otherwise require humans in the loop. Thus, cybersecurity becomes increasingly important in the case of CPSs, as all critical infrastructure is built using this model. The security of a CPS depends on more factors than conventional networked cyber systems. A secure CPS has at least the following features: confidentiality, integrity, availability, and authenticity. The various components of a cyber-physical system, along with points of vulnerability, are described in Figure 1.

FIGURE 1: REPRESENTATION OF A GENERIC CPS WITH ITS VARIOUS COMPONENTS AND POTENTIAL POINTS OF VULNERABILITY FOR AN ATTACK: A—PHYSICAL-CYBER ATTACK; B—PHYSICAL ATTACK; C, I, M—SUPPLY-CHAIN ATTACK, SIDE-CHANNEL ATTACK; D, H, L—FIRMWARE MODIFICATION ATTACK; F, J, N—NETWORK-BASED ATTACKS; E, G, K—CODE INJECTION, AND REPLAY ATTACK

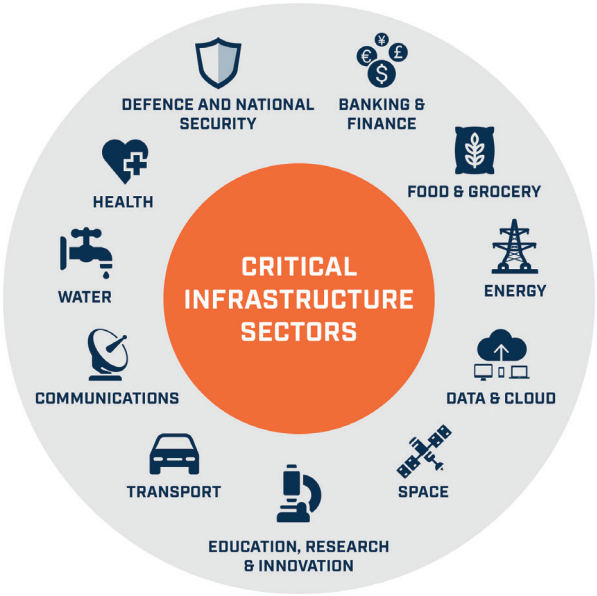


Thus, suffice to say, an insecure cyber system results in a vulnerable CPS. Furthermore, the risk is not limited to data and services but includes physical and biological hazards to life affected by the CPS. Since a physical process can be attacked physically by adversaries, the defense mechanisms need to consider physical security as well as cyber security. Thus, security preparedness is no longer restricted to IT; it is just as important to secure the OT domain.

B. Critical Infrastructure

It is stated in the Geneva Convention that the UN Security Council condemns attacks against critical civilian infrastructure and explicitly calls for the protection of this infrastructure during war. The Digital Geneva Convention in 2017 extended this concept further to digital public goods that have a cyber-physical impact [6]. Despite such efforts, rules for cyber warfare do not yet exist due to the nature of cyberspace, let alone rules for cyber-physical space. In such a scenario, a defender's advantage in the form of preparedness is paramount since all critical infrastructure sectors defined by the industrial control systems cyber emergency response team (ICS-CERT) in Figure 2 are possible targets.

FIGURE 2: CRITICAL INFRASTRUCTURE SECTORS DEFINED BY ICS-CERT

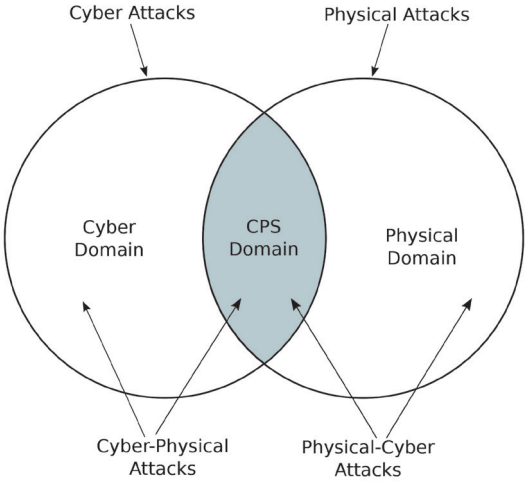


Due to its distributed nature, increased network connectivity, and technological readiness, modern national infrastructure [7] faces the following threats:

1. Cyberattacks—an attack exclusively in the cyber domain, where confidentiality, integrity, and availability are compromised.
2. Cyber-physical attacks—attacks in the cyber domain with consequences in the physical domain such as unauthorized actuation, espionage, and delayed actuation.
3. Physical attacks—attacks purely in the physical domain.
4. Physical cyberattacks—attacks on the physical domain with consequences in the cyber domain.

Cyber-physical attacks form the new frontier for CPXs. The relationships between the origins of attacks and consequential effects on different domains are illustrated in Figure 3, with the CPS domain being the most vulnerable of all the domains [8].

FIGURE 3: RELATIONSHIPS BETWEEN THE ORIGINS OF ATTACKS AND THE CONSEQUENCES OF SUCH ATTACKS ON THE VARIOUS DOMAINS



C. Cyber-Physical Exercise (CPX) versus Cyber Defense Exercise (CDX))

There is a need for security operators to familiarize their teams with knowledge of OT and industrial control systems. This knowledge includes cyber-physical security, safety, and servicing of critical infrastructure. Cyber-physical security is the set of tactics, techniques, and procedures (TTPs) that aid in the prevention of cyber-physical attacks. Physical effects can be wide-ranging, and an awareness of these impacts is necessary to estimate the stability of a system after an attack. Physical effects on a generic CPS may include delayed, wrong, unauthorized, or restricted actuation via cyber commands or physical force. Launching such an attack on actuators requires process knowledge beyond networking knowledge. This makes a CPX an ideal battleground for attackers (red teams), defenders (blue teams), evaluators (white teams), forensics (yellow team), and infrastructure builders (green teams). The scope of defense can also be increased to proactive, reactive, and detection-based defense.

Schoenmakers [9] remarked: "Differences in perspectives between IT and OT specialists can cause security issues for control systems. It is important for organizations to keep in mind that different values between groups can influence the perception of issues and solutions," which emphasizes the gap that exists between traditional IT security and ICS specialists. The knowledge gathered from such a CPX can help stakeholders understand the origin–impact relationship for different kinds of attacks [8] as described in Table II.

TABLE II: ORIGIN–IMPACT RELATIONS FOR ATTACKS [10]

Attack Type	Origin Domain	Impact Domain	Example
Cyber	Cyber	Cyber	Spear-phishing
Physical	Physical	Physical	Drilling a hole in a tank
Cyber-Physical	Cyber	Physical	Command injection
Physical-Cyber	Physical	Cyber	Signal jamming

3. CRITICAL INFRASTRUCTURE SECURITY SHOWDOWN: AN ANNUAL CPX

The Critical Infrastructure Security Showdown (CISS) is an annual CPX that utilizes operational critical infrastructure testbeds to serve as iTrust's technology assessment exercise. Dubbed the SWaT Security Showdown (S3) in 2016 and renamed S317 the following year, from 2019 onwards, the exercise came to be known as CISS [11], [12]. The one theme that remained consistent throughout these five years is "keep moving." Each year, the event had to adapt itself to the state-of-the-art in the cyber-physical security community.

The goals of this exercise are threefold: (a) to validate and assess the effectiveness of technologies developed by researchers associated with iTrust; (b) to develop capabilities for defending critical infrastructure under emergencies such as cyberattacks; (c) to understand composite tactics, techniques and procedures (TTPs) for enhanced operational security; and (d) to enable members of the red teams to understand operational approaches used to compromise critical infrastructure and hence identify the necessary protection mechanisms to defend against these approaches.

A. History

Until 2018, CISS was a week-long academic exercise called Secure Cyber-Physical Systems Week (SCyPhy Week) [11], [12]. In 2019, several international blue teams [13] from the industry were invited to evaluate their detection mechanisms. In 2020, it became the first event of its kind to be conducted entirely online with participants located remotely [14]. This motivated the need for telepresence technologies such as virtual/augmented reality, VPNs, and conferencing. In 2021, the cyber range included ransomware scenarios and a hybrid model of physical and digital twins for operational water treatment and distribution systems.

B. Critical Infrastructure: SWaT, WADI, and Digital Twin Testbeds

i. The Secure Water Treatment (SWaT) testbed consists of a modern six-stage process. The process begins by taking in raw water, adding necessary chemicals, filtering it using an ultrafiltration (UF) system, de-chlorinating it via ultraviolet lamps, and feeding it to a reverse osmosis (RO) system. A backwash process cleans the membranes in the UF system using the reject water from the RO system. The cyber portion of the SWaT testbed consists of a layered communications network, programmable logic controllers (PLCs), human–machine interfaces (HMIs) (as shown in Figure 6), a supervisory control and data acquisition (SCADA) workstation, and a historian system. Data from sensors is available to the SCADA system and recorded by the historian in a database for subsequent analysis. Figure 4 contains several pictures of the SWaT testbed.

FIGURE 4: (FROM LEFT TO RIGHT) WATER TANKS AT THREE DIFFERENT STAGES OF THE SWAT TESTBED, A FLOW METER SENSOR, AND AN ACTUATOR IN THE FORM OF A MOTORIZED VALVE

ii. The Water Distribution (WADI) testbed is a natural extension of the SWaT testbed, comprising two elevated reservoir tanks (see Figure 5), six consumer tanks, two raw water tanks, and a return tank. It is also equipped with chemical dosing systems, booster pumps and valves, instrumentation, and analyzers. The WADI testbed takes in a portion of the SWaT testbed's reverse osmosis permeate and raw water, thus forming a complete and realistic water treatment, storage, and distribution network. Integration of these testbeds enables researchers to experiment with the cascading effects of cyberattacks. In addition to attacks on and defense of the PLCs and networks, the WADI testbed also simulates the effects of physical attacks such as water leakage and malicious chemical injections. Unlike a water treatment plant, typically contained in a secured location, a distribution system comprises numerous pipelines spanning a large area, thus increasing the risk of physical attacks.

FIGURE 5: RESERVOIR TANKS AT DIFFERENT STAGES OF THE WADI TESTBED

iii. Digital Twins for the SWaT testbed

Having conducted and published research on the SWaT testbed, iTrust was invited to participate as a Green Team during Locked Shields 2021 and to contribute a special system in the form of a digital twin of the SWaT testbed. Built at iTrust by a group of researchers, the digital twin consists of six stages (shown in Figure 6) that mimic the behavior of the SWaT testbed at iTrust. For the Locked Shields 2021 exercise, it was later reduced to three stages (as shown in Figure 7), as explained in the following sections. The main task was to design and deploy this three-stage digital twin as a fictitious Berylian water purification plant (WP) and provide infrastructure support during the exercise. The WP was one of several special systems that each Blue Team was tasked to defend.

FIGURE 6: THE HUMAN–MACHINE INTERFACE OF THE DIGITAL TWIN OF THE SWAT TESTBED CLOSELY RESEMBLES THAT OF THE PHYSICAL TESTBED

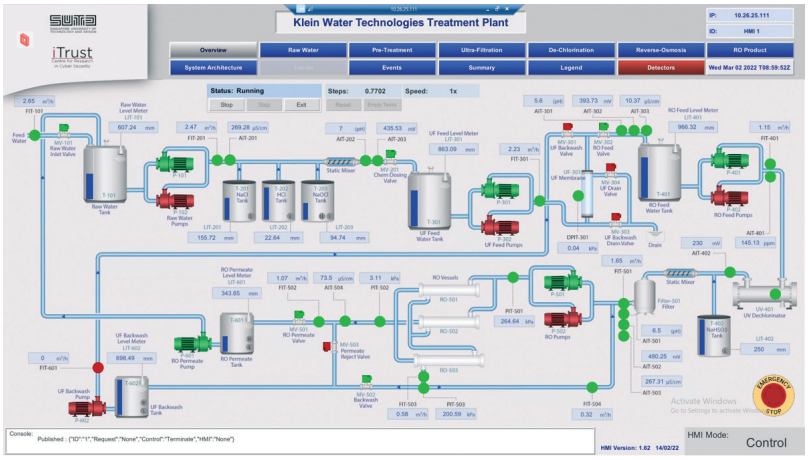
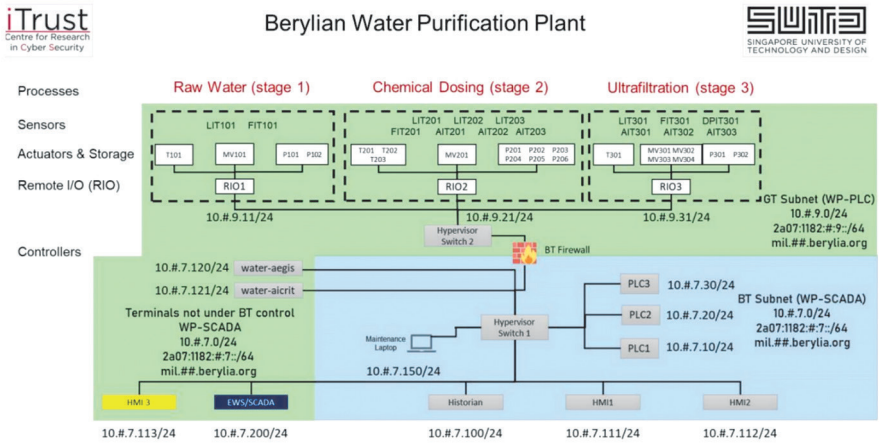


FIGURE 7: OVERVIEW OF THE THREE-STAGE DIGITAL TWIN OF THE SWAT TESTBED USED IN LOCKED SHIELDS 2021 AND CISS 2021



C. Participants

CISS was spread over two weeks and the aim was to meet the following objectives:

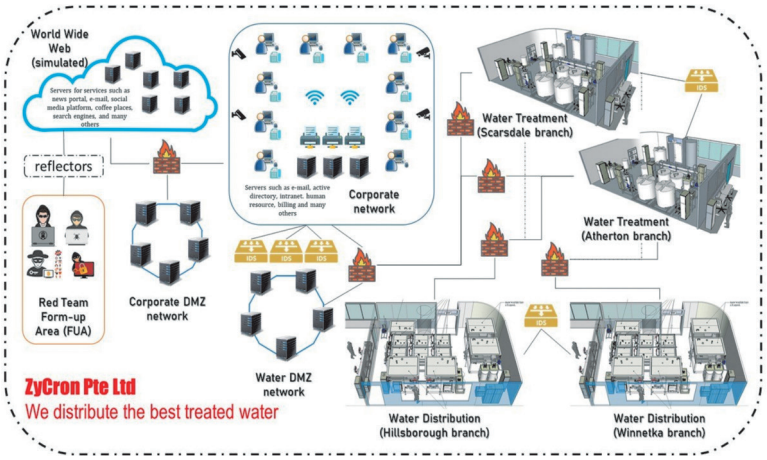
i. **Red Team** exercise: Each red team was given five hours to launch attacks on the platform. They would gain points when they met attack objectives. Blue team vendors were commercial companies, six of which had installed their products to detect attacks launched by the red teams. The TTPs were captured during this week.

ii. **Blue Team** exercise: A composite Red Team included multiple organizations. Attacks by the composite Red Team were defended by blue teams made up of critical information infrastructure (CII) operators and regulators (CII Blue Teams). Each CII Blue Team was given one eight-hour slot to respond to the attacks and defend the platform. Participation in the CII Blue Team was by invitation only.

iii. **Green Team** provided the infrastructure as shown in Figure 8. This involved an alert logger, an attack logger, plant visualizers, and anomaly detectors.

iv. **White Team** comprised the judges who were responsible for evaluating the performance of the blue and red teams. This is described in the following section (D).

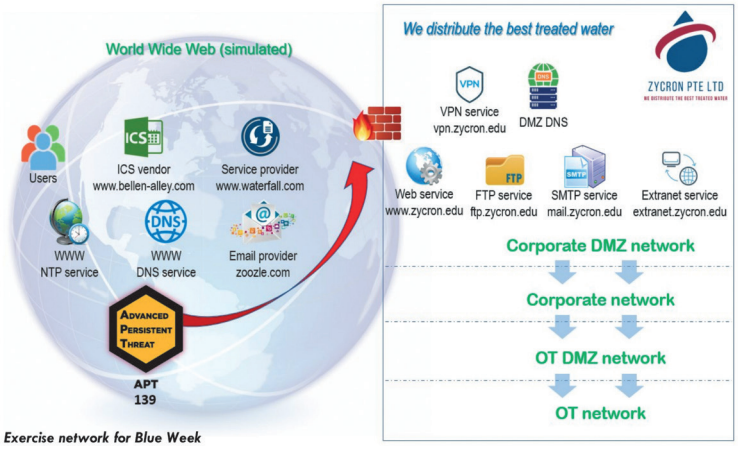
FIGURE 8: THE OVERALL CYBER-PHYSICAL RANGE IMPLEMENTED FOR THE CISS CPX



D. Scoring

The points awarded to the red teams for compromising various components are as follows: 100 points for the historian, PLCs, SCADA system, and network switches; 200 points for the flowmeter, water level meter, pressure meter, pumps, and valves; and 300 points for the conductivity meter, oxidation-reduction potential meter, and pH meter. The objective, attack method, tools, and description are noted for each attacker. Extra points are awarded for novelty. Figure 9 illustrates the points of compromise.

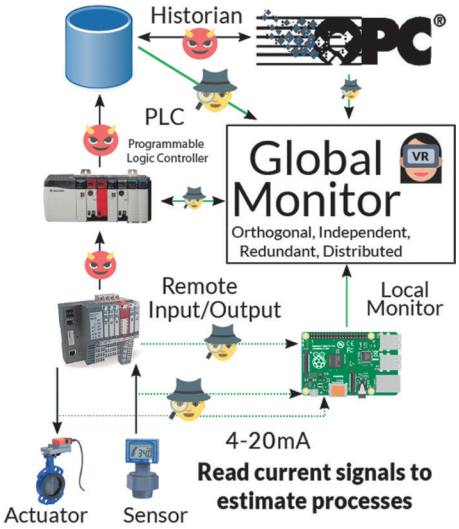
FIGURE 9: THE NUMBER OF NETWORK TRAVERSALS REQUIRED TO START EARNING POINTS IN DIFFERENT CATEGORIES



E. Attack Detection

Any alert generated by the security system deployed by a blue team was reported immediately to the plant operator—automatically, not manually. While each blue team was provided all event data, e.g., PCAP files and historian data, at the end of the event, they were not expected to analyze an alert generated during the event. Each alert was reported immediately, as if occurring in a live plant and being reported to the plant operator. This helped maintain credible neutrality in scoring and resulted in an accurate assessment of the results. Any anomaly resulting from the attack, or otherwise (e.g., a false alarm), and reported by one or more iTrust detectors was made visible only to the organizers, observers, and judges, but not to the red or blue teams. Figure 10 shows the locations of the various detectors.

FIGURE 10: DETECTORS MONITOR AS MANY POINTS OF COMPROMISE AS POSSIBLE TO AUTOMATE THE TASK OF SCORING AND TO EVALUATE THE DEFENSE TECHNOLOGY



The COVID-19 pandemic brought with it the challenge to conduct a CPX remotely with participants in multiple remote locations. Having the option of inviting remote participants opened up access to a broader set of attackers from around the world. CISS2021-OL (timeline provided in Table III) retained the online remote-first modality of CISS2020, with the following additions and changes:

i. The World Wide Web became the entry point.
ii. The Water Distribution (WADI) testbed was included as an additional attack surface.

iii. The number of hours for red teams to launch attacks was increased from four to five.
iv. Intrusion Detection Systems (IDSs) were installed.
v. A higher score was awarded to red teams that could avoid IDS detection.
vi. A higher score was awarded to red teams that used reflector servers to mask their identity.
vii. The prize money for the red teams was doubled.

TABLE III: A SAMPLE SCHEDULE THAT REPRESENTS THE DIFFERENT STAGES OF CISS FOR THE RED AND BLUE TEAMS

Phase	Date	Details	Involvement
Stage 1	5–16 Jul	Stage 1 (3 hours per team) to admit 10 Red Teams to the final round	Red Teams
	21 Jul	Red Team Finalists for CISS2021-OL were published	Red Teams
Online Briefing	23 Jul	Rules of engagement, attack objectives (Red Teams) Alert reporting (Blue Teams) Virtual tour of platforms (all) Q&A (all)	Red Teams
	26 Jul		CII Blue Teams
	30 Jul		Blue Team Vendoes
Red Team Preparation	30 Jul	Briefing for Red Team Finalists	Red Teams
	4–13 Aug	Familiarization: Testing VPN, RDP, attack tools	
	23 Aug	Submission of VMDK	
Blue Team Preparation	10 Aug	Blue Team Vendor Briefing	Blue Team Vendors
	16–31 Aug	Onsite deployment, baselining of products	Blue Team Vendors
Opening Ceremony	3 Sep	Welcome Address by organisers Tour of Testbeds	All
CISS2021 Finals	6–10 Sep	Red Team Exercise	Red Team, Blue Team Vendors
	13–17 Sep	Blue Team Exercise (closed door)	CII Blue Teams

4. PLAYBOOK FOR CYBER-PHYSICAL EXERCISES (CPXS)

Using the evidence-based playbook from the six years of CISS exercises from Section 3, we can generalize certain principles of good CPX design for different critical infrastructures.

A. Mission

The key goals/mission of a CPX should be firmly laid out. These serve as the anchor for all subsequent steps: (a) to validate and assess the effectiveness of technologies under development; (b) to develop capabilities for defending critical infrastructure during emergencies such as cyber-physical attacks; and (c) to understand composite tactics, techniques and procedures (TTPs) for enhanced operational security (OpSec).

B. Vision

i. To co-develop common and coordinated technical and strategic mobility against cyber-physical attacks that may occur on a national or international platform.
ii. To rapidly prototype continuity processes with cyber-physical security capabilities.
iii. To increase cooperation and coordination between the public and private sectors in the cyber-physical space.
iv. To gather evidence-based data from real-life scenarios.
v. To improve the maturity level of legal and regulatory compliance as well as technological and cultural acceptance for whole-of-society defense.

C. Timeline

The MITRE playbook can be extended to capture the uncertainty of cyber-physical systems. Unlike systems that are exclusively IT, these systems have mechanical/moving parts prone to frequent wear and tear. All timeline estimations should consider equipment failure, availability of backups, and fallback plans for each of the following phases:

i. Identifying: Concept development, initial planning phase
ii. Planning: Mid-term and final planning phase
iii. Conducting: Test run, main exercise
iv. Evaluating: Scoring, after action reports, reconnaissance

D. Tooling for First-Time Organizers and Participants

Maiden participation in any CPX requires significantly more effort because of a new

set of tools and infrastructure to learn. The following steps can help ease that journey:

i. Communicate and coordinate with the organizer
ii. Implement green team systems
iii. Implement networking protocols in the digital twin for IT and OT communication
iv. Distribute the digital twin across multiple VMs
v. Provide a Historian server and maintenance laptop for data storage
vi. Design and communicate the HMI: the human–machine interface
vii. Design and develop attacks for red team scenarios
viii. Implement anomaly detection for troubleshooting and technology evaluation
ix. Create test plans and update code according to the feedback received
x. Implement correctness testing of the twin and report suggested improvements

In addition to these, collaboration tools for file sharing, synchronous and asynchronous communication, code sharing, continuous integration and deployment, key management, and machine configuration must be in place for global coordination.

E. Evaluation and Incentive Design

Red teams are evaluated on stealth, persistence, impact, intention, success, and creativity. Blue teams get the opportunity to evaluate their commercial and academic products for anomaly detection. The Green Team sets up the entire infrastructure of the event to better prepare for participation in similar cyber exercises [13]. The White Team, comprising government, military, and plant operators, responds to incidents in real-time to improve their preparedness for anomalous scenarios.

F. Realism

Cyber-physical security is a new field. However, knowledge about these threats and how to respond to them is growing with each security incident. In 2021, there was an uptick in such incidents across power, water, nuclear, food, and oil/gas-based critical infrastructures. Therefore, it is important that attack scenarios in CPXs reflect these types of incidents, as illustrated in Table IV. To this end, future CPXs should also have ransomware injection attacks to prepare blue teams to help counteract such attacks using technology or military strategy.

TABLE IV: POPULAR CYBER-PHYSICAL ATTACKS THAT SHOULD BE REPRESENTED IN CPX SCENARIOS

Attack Type	Target	Origin	Cyber Impact	Physical Impact
Ukraine 2016 Spear-Phishing	Power Plant	HMI Malware	Loss of Availability	Blackouts
Dust Storm 2015 Watering Hole	Critical Infrastructures	Backdoor, Zero-Day	Loss of Confidentiality	None So Far
Stuxnet 2010	Centrifuge PLCs Code Injection	Commodity Computers	Loss of Integrity	Slowdown
Davids-Besse 2003 Slammer Worm	Nuclear Plant	Worm	Distributed DOS Attack	Plant Shutdown

G. Stress Tests

CPXs involve physical components which have a higher chance of failure. Therefore, stress tests under exercise conditions can be performed beforehand for added reliability. In the case of the SWaT digital twin, the following tests were conducted.

i. **Stress Test 1:**
Run the SWaT twin for one plant hour.
Execute the availability script once every minute.
Record the output of the script.
The test is rated as a pass if water is always available.

ii. **Stress Test 2:**
Run the SWaT twin for one plant hour.
Execute the availability script once every minute.
Launch each attack script three times at random times.
Record the output of each execution of the attack script.
The test is considered successful if the outcome of the script conforms to the intention of the attack.

iii. **Robustness Test (RT)**
Objective: To determine the robustness of the SWaT twin in the presence of environmental disturbances. Ideally, to be robust the SWaT twin must not crash or hang in the face of an environmental disturbance.

For example:
RT1: Stop one or more HMI processes and observe the twin's response.
RT2: Stop one or more RI processes and observe the twin's behavior.

RT3: Stop one or more PLC processes and observe the twin's behavior.
RT4: Stop one or more device processes, e.g. MV201, and observe the twin's behavior.

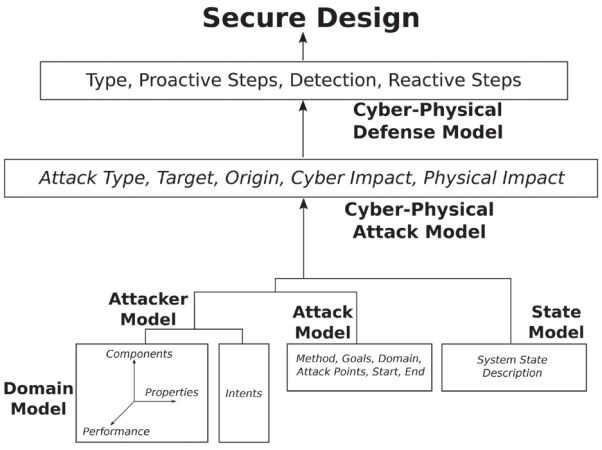
H. After Action Report

The data obtained from the exercise should cover all aspects of the NIST framework across the four stages described earlier. The framework is described in Figure 11.

FIGURE 11: NIST CYBERSECURITY FRAMEWORK APPLIED TO CPXS

An after action report is an important artifact that documents the security model for each of the five steps. These granular details are described in Figure 12.

FIGURE 12: THE AMOUNT OF GRANULAR INFORMATION THAT CAN BE OBTAINED FROM A SUCCESSFUL CPX



5. CONCLUSION

With the advent of cyber-physical systems, the fifth warzone of information warfare is now interdependent with land, air, water, and space. This increases the attack surface and necessitates high-quality CPXs to improve defense capabilities for defending national critical infrastructure. As CPXs become more popular, it is important to have an accessible and reproducible playbook for different organizers to conduct and participate in various relevant exercises. The composable and interoperable nature of the international security community hinges on the participatory interdependence of countries in each other's exercises. These CPXs serve as a battleground for ideas and invaluable resource sharing. In addition, they have broader international consequences in terms of security preparedness, responsible disclosure, timely triages of incidents, and an extended workforce in the form of allies during emergencies. Cyberspace offers a relatively cheap, fast, and easy domain to launch cyber-physical attacks. Furthermore, as yet, rules for cyber warfare are non-existent. Since cyberspace is inherently borderless, collaboration helps immensely in turning that feature into a strength. A reliable schedule of CPXs can establish a mechanism via which participating nations can share cyber-physical threats, indicators of compromise, forensic scenarios, targets, potential vulnerabilities, and timely advice. The quasi-competitive nature of CPXs also helps to promote a sense of camaraderie and excitement in the tightly knit security community to prepare the participants well for any potential real-world scenarios that involve cyber-physical conflict and cooperation.

ACKNOWLEDGMENTS

Several iTrust staff members contributed to this paper by helping to organize CISS over six years, by participating in international CPXs, by developing cyber-physical infrastructure to facilitate the exercises, by managing teams of people, by reviewing after-action reports, or by recording the playbooks to use in reports such as this one. These include Andrew Tay, Angie Ng, Desmond Wan, Ian Teo, Ivan Lee, Kaung Myat Aung, Mavis Ting, Priscilla Pang, Salimah Liyakkathali, Sridhar Adepu, and others. Thank you to everyone who took part in ensuring this paper's accuracy and completeness. This work is supported (in part) by the National Research Foundation, Singapore, and the Cyber Security Agency of Singapore, under its National Cybersecurity R&D Programme (NRF2018NCRNSOE005-0001). Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors. They do not necessarily reflect the views of the National Research Foundation, Singapore, and the Cyber Security Agency of Singapore.

REFERENCES

[1] The World Economic Forum, "Global Cybersecurity Outlook 2022," Insight Report, Jan. 2022.

[2] C. Fradkin, "Cyberattacks in 2021 highlighted critical infrastructure risks," Security Boulevard. Accessed: Jan 2, 2022. [Online]. Available: https://securityboulevard.com/2021/11/cyberattacks-in-2021-highlighted-critical-infrastructure-risks/

[3] E. Seker and H. H. Ozbenli, "The Concept of Cyber Defence Exercises (CDX): Planning, Execution, Evaluation," in *2018 Int. Conf. Cyber Secur. Prot. Digit. Serv., Cyber Secur*, 2018, pp. 1–9, doi: 10.1109/CyberSecPODS.2018.8560673.

[4] "Locked Shields Exercise Featured a Connection between Cyber and Information Operations." CCDCOE. https://ccdcoe.org/news/2021/locked-shields-exercise-featured-a-connection-between-cyber-and-information-operations/ (accessed Dec. 7, 2021).

[5] J. Kick, "Cyber exercise playbook," MITRE Corp Bedford, MA, USA, 2014.

[6] V. Jeutner, "The Digital Geneva Convention: A Critical Appraisal of Microsoft's Proposal," *J. Int. Humanit. Leg. Stud.*, vol. 10, no. 1, pp. 158–170, 2019.

[7] J. Brynielsson, U. Franke, M. A. Tariq, and S. Varga, "Using cyber defense exercises to obtain additional data for attacker profiling," in *2016 IEEE Int. Conf. Intell. Secur. Inform., ISI* 2016, pp. 37–42, doi: 10.1109/ISI.2016.7745440.

[8] F. Skopik and M. Leitner, "Preparing for National Cyber Crises Using Non-linear Cyber Exercises," in *2021 18th Int. Conf. Priv., Secur. Trust, PST* 2021, pp. 1–5, doi: 10.1109/PST52912.2021.9647795.

[9] F. A. Schoenmakers, "Contradicting paradigms of control systems security: how fundamental differences cause conflicts," 2013.

[10] Loukas G. Cyber-physical attacks: A growing invisible threat. Butterworth-Heinemann; 2015 May 21.

[11] D. Antonioli, H. R. Ghaeini, S. Adepu, M. Ochoa, and N. O. Tippenhauer, "Gamifying ICS security training and research: Design, implementation, and results of s3." *CPS-Sec Int. Workhop Cyber-Phys. Sys. Secur*. 2017, New York, NY, USA, 2017, pp. 93–102.

[12] "S3-2016: SWaT Security Showdown (S3)." iTrust. https://itrust.sutd.edu.sg/scy-phy-systems-week/2016/s3/ (accessed Dec. 13, 2021).

[13] "CISS: Critical Infrastructure Security Showdown 2020-OL." iTrust. https://itrust.sutd.edu.sg/ciss/ciss-2020-ol/ (accessed Dec. 1, 2021).

[14] "CISS: Critical Infrastructure Security Showdown 2019." iTrust. https://itrust.sutd.edu.sg/ciss/ciss-2019/ (accessed Dec. 10, 2021).

[15] "CISS: Critical Infrastructure Security Showdown 2021." iTrust. https://itrust.sutd.edu.sg/ciss/ciss-2021/ (accessed Dec. 13, 2021).

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

Data Quality Problem in AI-Based Network Intrusion Detection Systems Studies and a Solution Proposal

Maj. Emre Halisdemir
Strategy Researcher
CCDCOE
Tallinn, Estonia
Emre.Halisdemir@ccdcoe.org

Hacer Karacan
Associate Professor
Gazi University
Ankara, Turkey
hkaracan@gazi.edu.tr

Mauno Pihelgas
Technology Researcher
CCDCOE
Tallinn, Estonia
Mauno.Pihelgas@ccdcoe.org

Toomas Lepik
Information Security Expert
Tallinn University of Technology
Tallinn, Estonia
Toomas.lepik@taltech.ee

Sungbaek Cho
Strategy Researcher
CCDCOE
Tallinn, Estonia
Sungbaek.Cho@ccdcoe.org

Abstract: Network Intrusion Detection Systems (IDSs) have been used to increase the level of network security for many years. The main purpose of such systems is to detect and block malicious activity in the network traffic. Researchers have been improving the performance of IDS technology for decades by applying various machine-learning techniques. From the perspective of academia, obtaining a quality dataset (i.e. a sufficient amount of captured network packets that contain both malicious and normal traffic) to support machine learning approaches has always been a challenge. There are many datasets publicly available for research purposes, including NSL-KDD, KDDCUP 99, CICIDS 2017 and UNSWNB15.

However, these datasets are becoming obsolete over time and may no longer be adequate or valid to model and validate IDSs against state-of-the-art attack techniques. As attack techniques are continuously evolving, datasets used to develop and test IDSs also need to be kept up to date. Proven performance of an IDS tested on old attack patterns does not necessarily mean it will perform well against new patterns. Moreover, existing datasets may lack certain data fields or attributes necessary to analyse some of the new attack techniques.

In this paper, we argue that academia needs up-to-date high-quality datasets. We compare publicly available datasets and suggest a way to provide up-to-date high-quality datasets for researchers and the security industry. The proposed solution is to utilize the network traffic captured from the Locked Shields exercise, one of the world's largest live-fire international cyber defence exercises held annually by the NATO CCDCOE. During this three-day exercise, red team members consisting of dozens of white hackers selected by the governments of over 20 participating countries attempt to infiltrate the networks of over 20 blue teams, who are tasked to defend a fictional country called Berylia.

After the exercise, network packets captured from each blue team's network are handed over to each team. However, the countries are not willing to disclose the packet capture (PCAP) files to the public since these files contain specific information that could reveal how a particular nation might react to certain types of cyberattacks. To overcome this problem, we propose to create a dedicated virtual team, capture all the traffic from this team's network, and disclose it to the public so that academia can use it for unclassified research and studies. In this way, the organizers of Locked Shields can effectively contribute to the advancement of future artificial intelligence (AI) enabled security solutions by providing annual datasets of up-to-date attack patterns.

Keywords: *intrusion detection systems, datasets, artificial intelligence, Locked Shields*

1. INTRODUCTION

IDS development is an important cybersecurity area that has been studied for many years. The first IDS studies started in the 1980s utilizing expert systems [1]. In the years since, many machine learning and AI methods have been used in IDS development studies [2]–[13]. In the literature review, we saw that deep learning, a type of AI technique, is mostly used in recent IDS development studies [14]–[17].

The models used for efficient IDS solutions are very important. In addition, the quality of the datasets used for training and testing the models is also very important in IDS development studies. There are many features that can reveal the quality of an IDS dataset, such as network traffic generation time, proportion of normal and attack traffic, amount of data, whether it is generated from a real or emulated network, type of network (corporate network, small network, university network, military network etc.), and label type. The quality of the dataset paves the way for the development of better-trained models and therefore better intrusion detection.

Several publicly available datasets are used in IDS development. These datasets are not only applicable to AI-centric applications but are also suitable for traditional IDS development studies. In addition, governmental institutions also carry out IDS dataset generation studies. However, many institutions cannot make their datasets public due to security and privacy restrictions. Thus, IDS development studies are often carried out using highly limited public datasets.

In this study, firstly, we overview the production process of the datasets used in IDS development studies. Then, we examine the characteristics of commonly used public datasets, discuss which dataset can be effective in detecting which attack types, and compare each dataset in terms of its advantages and disadvantages. In the following section, we introduce the Locked Shields (LS) exercise and analyse the advantages of using the PCAP files obtained from the exercise. Finally, we identify the obstacle to disclosing these files to the public and suggest a solution to solve this problem.

2. INTRUSION DETECTION SYSTEM DATASETS PRODUCTION

There are two important aspects of IDS development studies. One of them is model development and the other is dataset quality. Model development studies have evolved into AI-based techniques and the results have proven their success. On the other hand, generating datasets of an appropriate quality and quantity of data remains a problem [18]. Moreover, there are problems in the evaluation, comparison, and implementation of IDSs, mainly because of the lack of appropriate datasets. Most organizational datasets are not publicly available due to security and privacy concerns, and public ones do not reflect current attack trends.

The datasets produced by MIT Lincoln Lab at the request of the US Defense Advanced Research Projects Agency (DARPA) and the KDD datasets generated by the University of California contributed greatly to IDS studies for many years. However, accuracy problems and the fact they do not reflect real situations have caused criticism [19], [20].

On the other hand, as network behaviour patterns change and intrusion methods evolve, it becomes necessary to shift from the use of static and single-use datasets to dynamic datasets that reflect up-to-date traffic composition and intrusion methods. These datasets should also be modifiable, expandable, and reproducible.

The datasets used in network intrusion detection are typically produced by monitoring the network traffic and processing the traffic data. There are two approaches to capturing network traffic: packet-based and flow-based. Packet-based capture is conducted by mirroring the traffic on network devices to a network capture node. The captured data includes all network header and payload information. Packet-based network traffic data is usually stored in packet capture (PCAP) format. The characteristics of the captured data depend highly on the type and purpose of the mirrored network. For example, the ratios between various network protocols such as Transmission Control Protocol (TCP), User Datagram Protocol (UDP), Internet Control Message Protocol (ICMP) and Internet Protocol (IP) can vary significantly between different networks. Furthermore, the type of payload content, the amount of encrypted traffic, and the presence of normal and malicious activity depends entirely on the capture scenario and situation. In contrast, flow-based capture provides more aggregate data and includes metadata about network connections. The main flow-based formats are NetFlow, IPFIX, sFlow, and OpenFlow. It is possible to convert packet-based data to flow-based data via applications like nfdump and Yet Another Framework (YAF). In addition to packet-based and flow-based datasets, there are also datasets such as KDDCUP99, which was generated by enriching flow-based datasets with packet-based or host-based data [21].

The production methods of and information contained in IDS datasets differ from one another. Furthermore, some datasets have been specially developed for specific purposes in IDS studies. Thus, each dataset offers a different set of characteristics, which makes it difficult to compare individual datasets. However, there are several properties to consider when evaluating a specific IDS dataset, as shown in Table I.

TABLE I: VALUE RANGES OF IDS DATASET PROPERTIES [22]

General Information	**Traffic generation time**	Day/Month/Year
	Public	No / no info / if requested / yes
	Normal traffic	No / yes
	Attack traffic	No / not applicable (NA) / yes
Nature of Dataset	**Metadata**	No / some / yes
	Format	Bidirectional flow / logs / others / packet / one direction flow
	Anonymity	No / NA / yes / yes for some features
Data Quantity	**Quantity**	Data amount in GB / flow / packet / point numbers
	Time	Data save time
Recording Environment	**Traffic type**	Emulated / real / synthetic
	Network type	Different networks / corporate network / honeypots / Internet service provider / NA / small network / military network / university network
	Full network	No / NA / yes
Evaluation	**Previously defined sections**	No / NA / yes
	Balanced	No / NA / yes
	Labelled	Relatively / no / yes / yes (IDS) / yes with the background

Each field in Table I provides information about the properties of a dataset. For instance, the traffic generation time indicates whether or not current attack types are included. Another important feature of a dataset is that it contains normal traffic. An IDS dataset must contain both attack and normal traffic. Thus, if a dataset does not contain normal traffic, it can only be used by combining it with another dataset or actual network traffic. Format information gives an idea about the dataset creation process. Quantity information relates to the size of a dataset and this can be used to evaluate whether the system can process the dataset. Consequently, this information as a whole can help to determine the quality of a dataset and be used to evaluate the dataset.

3. EXAMINATION OF COMMONLY USED IDS DATASETS

In this section, we examine a selection of six public datasets that are frequently used in IDS development studies. These datasets were produced in different environments using different methods and each dataset has its own specific features. After clarifying

the objectives of the IDS to be developed, the dataset features should be examined, and the appropriate dataset selected to train and test the model.

The oldest of this selection is the KDDCUP99 dataset obtained from US Air Force network traffic. Published in 1999, this dataset served IDS development studies for many years. However, problems in the dataset prompted the need for a new dataset. Hence, the NSL-KDD dataset was published in 2009. The UNSW-NB15 dataset produced by the Australian Cyber Security Centre using the IXIA PerfectStorm tool was subsequently published in 2015. One of the most widely used datasets in recent years, the UGR'16 dataset was generated from data sourced from the network of a Spanish information service provider (ISP) in 2016. The last two datasets were published in 2017: the CIDDS-001 dataset was generated by Coburg University of Applied Sciences in Germany and the CICIDS 2017 dataset was generated by the Canadian Institute for Cybersecurity. The characteristics of each of these datasets are summarized in Table II.

TABLE II: FEATURES OF COMMONLY USED IDS DATASETS

	General Information						Data size		Environment	
Dataset	**Traffic Gen. Year**	**Public**	**Normal Traffic**	**Attack Traffic**	**Format**	**Size**	**Duration**	**Traffic Type**	**Network Type**	**Label-led**
CICIDS 2017	2017	Yes	Yes	Yes	Bi-direc-tional flow	51.3 GB	5 days	Emu-lated	Small network	Yes
CIDDS-001	2017	Yes	Yes	Yes	One-direc-tion flow	380 MB	28 days	Emu-lated and real	Small network	Yes
UGR'16	2016	Yes	Yes	Yes	Bi-direc-tional flow	~17 GB	4 months	Real	ISP	Yes
UNSW-NB15	2015	Yes	Yes	Yes	Packet and others	2 MB	31 hours	Emu-lated	Small network	Yes
NSL-KDD	2009	Yes	Yes	Yes	Other	150 KB	N/A	Emu-lated	Small network	Yes
KDD CUP99	1999	Yes	Yes	Yes	Other	5 MB	N/A	Emu-lated	Small network	Yes

As shown in Table II, some of the datasets are quite old and therefore do not include up-to-date attack types. An effective dataset should contain realistic and up-to-date attacks to improve the detection capability of an IDS. All of the datasets presented in the table are public and contain normal and attack traffic and thus are suitable for machine learning. Including a fair amount of normal traffic is important for dataset

quality, and more realistic background traffic provides a more effective dataset. As for packet capturing formats, each dataset has been generated using different formats. It is generally accepted that a quality dataset should include complete traffic and therefore bi-directional datasets are usually preferred for better performance.

In terms of size, CICIDS-001 is the largest dataset and NSL-KDD is the smallest. Although it is important to have good quality and a sufficient amount of data in the dataset, IDS models may have difficulty processing extremely large datasets. For this reason, it is important to produce an optimal size dataset containing a variety of records from all attack types along with normal traffic.

Data capture duration also varies by dataset and this information is not accessible for some datasets. Duration data is important, especially for datasets generated from real networks, because it provides information on the time periods from which a dataset was produced. With this information, a detailed analysis of how well the dataset captured malicious network traffic can be conducted.

Network traffic is generally generated from emulated networks because a variety of attack types can be synthesized and included in the datasets. The UGR'16 dataset contains real network traffic as it was obtained from the internet service provider (ISP) network. A balance of synthetic and realistic data is important to create a quality dataset. All datasets contain labelled data and it is vital that the records in the dataset are correctly labelled as malicious or benign. In case of incorrect labelling, the model will be trained and tested incorrectly, leading to problems in attack detection in real networks.

When evaluating a dataset, it is essential to know the types of attacks it contains. If the models are not trained with datasets containing the appropriate attack types, it will be impossible to obtain adequate results. Therefore, the relationship between attack types and datasets was examined within the scope of the study. The results are shown in Table III.

TABLE III: IDS DATASETS AND ATTACK TYPES

Dataset	Attack Categories
CICIDS 2017 [23]	JCross-site-scripting botnet, SQL injection, DoS, DDoS, Heartbleed, infiltration, SSH brute force
CIDDS-001 [24]	SSH brute force, DoS, port scans
UGR'16 [25]	Spam, botnet, DoS, port scans, SSH brute force
UNSW-NB15 [26]	DoS, backdoors, generic, exploits, fuzzers, worms, port scans, reconnaissance, shellcode, spam
NSL-KDD [27]	DoS, probing, remote-to-local, user-to-root
KDD CUP99 [28]	DoS, probing, remote-to-local, user-to-root

The table lists the six most popular datasets used in IDS development and the attack categories they contain. UNSW-NB15 has the largest number of attack categories, in contrast to CIDDS-001, which has the smallest number of attack categories. However, the CIDDS-001 dataset contains detailed metadata that can be used for further analysis [24]. NSL-KDD contains the same attack categories as KDD CUP99 as it is generated from the KDD CUP99 dataset by removing redundant records from the KDD CUP99 dataset and increasing the difficulty level of the records [27]. The attack categories in the UGR'16 dataset were derived from the real traffic of an ISP. Some of the attacks shown in the table were obtained from real traffic, and some were artificially synthesized attacks [25].

Regardless of their different properties, such as format, size, and attack types, each dataset has uniquely contributed to the development of IDSs based on machine learning. However, there are also some criticisms of each dataset. Although the KDD CUP99 dataset has made significant contributions to IDS development studies in the past, it has been strongly criticized due to its low level of difficulty and the presence of redundant records. It was revised in 2009 to create the NSL-KDD dataset, which is still frequently used in IDS development studies. However, NSL-KDD does not have a normal traffic distribution due to the small number of records it contains for some attack types [27]. The UNSW-NB15 dataset was produced in a short period of 31 hours using the traffic generator IXIA Perfect Storm. It includes attacks from many attack categories. From the perspective of attack types and dataset size, it can be used effectively in IDS development studies. A disadvantage, however, is that it was produced from a small emulated network [26].

The most important feature of the UGR'16 dataset is that it was generated from a real

ISP network, produced over a period of four months. However, the dataset, which contains both normal and attack traffic, is inevitably quite large compared to other flow-based datasets [25]. The CIDDS-001 dataset was produced over a period of four weeks and contains metadata for detailed analysis. In addition, scripts that enable the processing of normal and malicious user behaviour have been published [29]. It was, however, generated from an emulated small business environment and contains fewer attack types than other datasets. Currently, the CICIDS 2017 dataset is the most advantageous in that it is the most recent of these public datasets. But it is already over five years old and the fact it was produced using a synthetic mesh is a drawback.

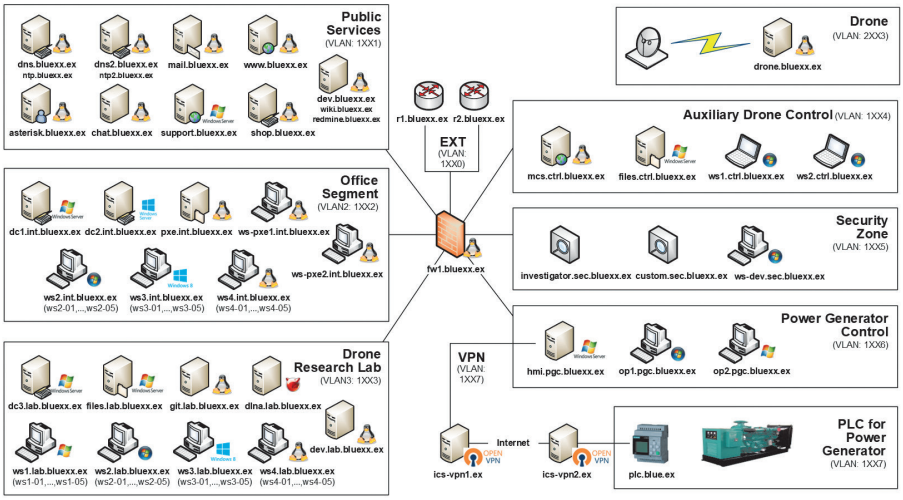
4. DATA FROM THE LOCKED SHIELDS EXERCISE

Locked Shields (LS) is an annual live-fire cybersecurity exercise that has been organized by the NATO CCDCOE since 2010. In the exercise, blue teams learn and test their skills in many interdisciplinary categories, for instance, defending against real-time cyberattacks, situation assessment, incident response, handling various inject scenarios, and maintaining the functionality of their computer systems.

As LS is a defence-oriented exercise, the primary training audience is the blue teams. Other exercise participants belong to one of four supplementary teams: the red team conducts attacks; the yellow team provides situation awareness; the green team upholds the backend infrastructure; and the multifaceted white team provides exercise control, strategy game, legal game, media etc. Blue teams fulfil the role of rapid-reaction teams dispatched to assist a fictitious country, Berylia, which is in a prolonged conflict with another fictional country, Crimsonia, primarily represented by the red team.

Blue teams are required to defend a variety of typical IT systems as well as specialized industrial control systems – for instance, Linux and Windows servers, Linux and Windows workstations, FreeBSD firewalls, industrial programmable logic controllers (PLCs), professional power management systems, water treatment plants, air defence systems, and 4G/LTE gateways. Figure 1 illustrates an example network that the blue teams must defend during the exercise.

FIGURE 1: EXAMPLE NETWORK MAP OF THE LOCKED SHIELDS EXERCISE



Blue teams undertake system administration and hardening techniques, forensic and legal challenges, as well as handling various other injects from the white team. Consequently, blue team participants must be experts with a variety of skills to cover all necessary competencies. However, the defending teams must consider that there is also a dedicated user simulation sub-team that assumes the role of various users working on the systems. In addition to mimicking normal usage patterns, they also insist blue teams must keep their systems functional. Although preparation long precedes the main event, the intense live-fire gameplay of the exercise unfolds over just two days.

The red team conducts a variety of escalating attacks – from initial access, continuing persistence, and privilege escalation to data collection, exfiltration, and destruction. Due to the short time span of the exercise, the number and pace of attacks are high compared to any real environment. Attack techniques (according to the MITRE ATT&CK knowledge base classification) from prior exercises include exploiting public-facing applications, compromising valid user accounts, exploiting privilege escalation, lateral movement using remote service, data collection and exfiltration from target systems, defacement of websites, and denial of service.

A. Exercise Data Capture

While capturing data during cyber exercises is a common practice [30]–[32], public release of these exercise datasets (e.g. the Cyber Czech dataset [33]) is less common.

The network traffic from NATO CCDCOE technical cybersecurity exercises is typically captured and stored for later analysis. The input data feed from the NATO Cyber Range is provided as ERSPAN (Encapsulated Remote Switched Port Analyser) mirror sessions from the switches in the exercise network environment. As a result, the data capture includes traffic from all (including internal) networks segments, not just traffic between the routers and perimeter firewalls.

NetFlow data from some routers would technically be available, but it was decided not to use it because it would only provide limited visibility compared to the full packet capture from all networks. In addition, PCAP files can be later processed and transformed into NetFlow data. Furthermore, IDS/IPS solutions can be configured to log textual network flow information that is very similar to the NetFlow format.

PCAP data collected during the exercise are split into separate files for each blue team. These blue-team-specific PCAP files are then shared with the corresponding nation after the exercise has concluded. Therefore, nations can typically analyse their PCAP files to conduct an after-action assessment, improve their defensive techniques, develop novel research [34], and prepare for the next exercise. However, this PCAP data may include information that can be used to determine a nation's defensive tactics, response patterns, and toolsets utilized to defend against various cyberattacks. Therefore, as a precaution, these PCAP files are considered sensitive and cannot be made public.

Furthermore, it is difficult to ensure no other sensitive data is included in the full packet capture data. Inclusion of personally identifiable information (e.g. names, usernames, e-mail addresses) of participants can occur unintentionally, for instance, if a participant were to use a personal device to connect to the exercise environment.

Creating a simple unoccupied blue team has already been attempted in past exercise iterations; however, without actual players, this empty team did not achieve the same level of interaction with the red and user simulation teams as the other blue teams. Thus, actual players within a blue team are necessary to create a more realistic dataset.

To solve this issue, we introduce the concept of the *research team*. This team will operate similarly to other teams; however, data captured from this team will be prepared for public release. It is possible to sanitize PCAP files using tools such as tcprewrite, TracerWrangler and Bit-Twist; however, we aim to keep such post-mortem modifications to a minimum. Regardless of whether such a tool is used, it is essential that any changes to MAC and IP addresses are coherent and deterministic, so that an individual specific IP is always transformed into the same target IP for all sessions. Otherwise, the sanitized dataset would become meaningless.

Initially, the research team in the LS exercise will consist of volunteer players who fulfil all the necessary roles like a normal team. With the recent advances toward autonomous cyber defence [35], [36], the future research team should ideally be a virtual team, or at least a hybrid between an AI cyber defence component and a small number of human players. Although virtual players in cybersecurity exercises have been experimented with in the past (e.g. during the 2016 DARPA Cyber Grand Challenge [37]), the environment and the scenarios were significantly more constrained compared to the LS exercise.

The virtual team would be deployed as customized intelligent cyber defence agents that would employ both signature- and anomaly-based detection and defence techniques (such as the architecture described in [35]). The agents can monitor themselves as well as the systems they are designated to defend and preserve. The agent must be aware of the surrounding conditions and the purpose of the host system because defences can vary significantly between servers, workstations, and industrial control systems. Furthermore, the agents should form a cluster and share data (e.g. detection patterns, indicators of compromise) among themselves. After planning, the agent can take actions such as blocking connections, reconfiguring firewall rules, restarting service modules, and sending notifications to other agents and human cyber defenders.

B. Producing a Publicly Shareable Dataset

Sharing PCAP files for any purpose can be challenging due to the large file sizes. The entire LS full packet capture produces approximately 11 TB of PCAP data for the two days of the exercise. Furthermore, over the years, this number has been slowly but steadily growing due to the increasing number of participating blue teams, the expansion of the exercise environment, and the overall size of the deployed systems. When this PCAP data is split up into individual blue teams, the median size of the data for a single team is about 400 GB. File compression typically provides approximately 40% space saving, resulting in about 240 GB of compressed PCAP data per team. Note that these are approximations based on an assessment of LS data capture since 2017. The actual data size can vary significantly from one team to another due to various in-game actions (e.g. large downloads and uploads, additional logging, service proxying) taken by the defending blue teams. Due to the sensitivity of team-specific data, the files are made available only to the corresponding blue teams over an encrypted channel.

The proposed research team will likely produce a similar amount of network data. This would mean an exceedingly large amount of raw data from various networks of mixed types. Simply providing this large amount of data would not be enough, as it would result in a dataset that is difficult for researchers and analysts to manage and

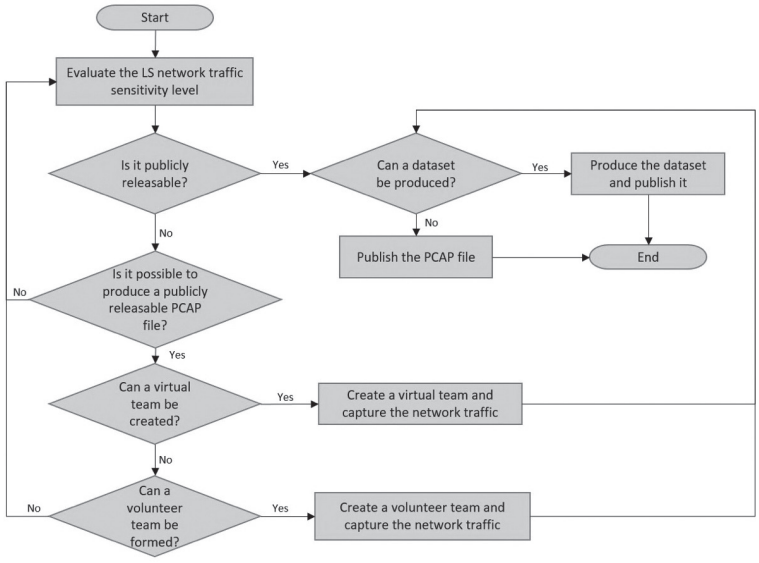
comprehend. Therefore, the published data needs to be more manageable for future research endeavours.

We propose to split or filter the data once more by the captured network type. For example, according to Figure 1, the data could be split into different collections based on the defined network segments, such as Public Services, Office Segment, Drone Research Lab, and Power Generator network. This would result in several smaller PCAP collections that are less complicated to handle. Since IDS configuration and network defence techniques can differ significantly across public, private, and special-purpose networks, this kind of separation would also benefit detection algorithm development.

Moreover, the overall value will increase by producing an annual LS dataset with enriched network flow information. The data will be labelled using a combination of an IDS software solution as well as the input from the red team. The flowchart representing the generation process of a unique and novel dataset from the LS environment is shown in Figure 2.

The proposal we present in this article should be carried out in coordination with the LS exercise organization team. For this reason, we discussed the applicability of our proposal with the exercise coordinator and team leaders. Our experiments continued with their positive feedback and support.

FIGURE 2: FLOWCHART REPRESENTING THE PROPOSED UNCLASSIFIED IDS DATASET GENERATION PROCESS FROM THE LOCKED SHIELDS EXERCISE



5. IMPLICATIONS OF THE STUDY

In this paper, we examined the properties of six datasets that have been used most frequently in IDS development studies. Its aim is to analyse the advantages and disadvantages of a dataset to be produced from the LS exercise. Similar to Table II, which summarizes the properties of the selected existing datasets, the LS dataset properties are summarized in Table IV. In addition to the information specified in Table IV, metadata can also be produced for the dataset. To create an efficient dataset for research, decisions on which sub-network to choose from the entire exercise infrastructure and how to adjust the attack and normal traffic balance should be made in a sound manner. The anonymity of the data will also need to be evaluated thoroughly prior to publication.

TABLE IV: LOCKED SHIELDS IDS DATASET FEATURES

	General Information						Data size		Environment	
Dataset	**Traffic Gen. Year**	**Public**	**Normal Traffic**	**Attack Traffic**	**Format**	**Amount**	**Duration**	**Traffic Type**	**Net-work Type**	**Label-led**
LS (Year)	Will be updated annually	TBD	Yes	Yes	Packet	TBD	2 + TBD prepa-rations days	Real + synt-hetic	Diffe-rent net-works	Yes (IDS)

Another important feature of an IDS dataset is the attack diversity it contains. Table III shows the attack categories in the six datasets. The attack types included in the LS exercise are determined by the red team based on their research carried out every year. Within the scope of the exercise, up-to-date attack types are applied every year and the responses from the blue teams to these attacks are measured. To compare the attack categories that the LS dataset may provide with those in the six evaluated datasets, the attack categories used by the LS exercises in past years are presented in Table V.

TABLE V: CATEGORIES OF ATTACKS CONDUCTED IN PAST LOCKED SHIELDS EXERCISES

Dataset	Attack Categories
Locked Shields	Exploit public-facing applications, compromise valid user accounts, exploit for privilege escalation, lateral movement using remote services, data collection and exfiltration from target systems, defacement of websites, denial of service

Conducting the exercise every year will ensure that up-to-date attack categories are added and up-to-date infrastructure devices are used. In addition, the fact that the exercise infrastructure is quite extensive will offer several advantages for producing a realistic dataset. There are many different systems with specific network infrastructure in the exercise. Addressing some of these each year will also provide an important opportunity to observe the network behaviour of different environments. Furthermore, background traffic generation in the exercise will contribute to the normal traffic distribution. Therefore, the dataset will include a balance of normal and attack traffic. As a result, considering the dataset features and the attack categories, it can be expected that a dataset created from the unique infrastructure of the LS exercise will make a significant contribution to future IDS development studies carried out in academia and industry.

6. CONCLUSION AND FUTURE WORK

In this paper, we analysed the process of generating IDS datasets and examined the characteristics of six commonly used public datasets. We introduced the annual Locked Shields exercise, examined the advantages of a dataset generated from the exercise traffic, and proposed a method to create a publicly releasable dataset. Academia and industry may benefit from this approach to obtain a stable source for IDS datasets, as the LS exercise would provide up-to-date attack patterns annually.

These are the first steps towards generating a publicly available dataset from the LS exercise. Subsequent considerations to be made in the future include: integrating a team of human volunteers into the LS exercise; capturing the network traffic of this team; pre-processing the obtained network traffic data; generating datasets from the PCAP file using various techniques; executing virtual team building activities; integrating the virtual team into the LS exercise infrastructure; and capturing network traffic from the virtual team via autonomous techniques.

REFERENCES

[1] J. P. Anderson, 'Computer Security Threat Monitoring and Surveillance', Technical Report, James P. Anderson Company, Fort Washington, PA, USA, 1980.

[2] J. Zhang, G. Qin, Y. Cui, J. Dong, and L. Guo, 'SVM-FastICA Based Detection Ensemble System of EEG', presented at the International Conference on Convergence Information Technology, Gyeongju, South Korea, 21–23 November 2007, pp. 2247–2253.

[3] J. Yuan, H. Li, S. Ding, and L. Cao, 'Intrusion Detection Model based on Improved Support Vector Machine', presented at the Third International Symposium on Intelligent Information Technology and Security Informatics, Jinggangshan, China, 2–4 April 2010, pp. 465–469.

[4] J. Jiang, R. Li, T. Zheng, F. Su, and H. Li, 'A new intrusion detection system using Class and Sample Weighted C-Support Vector Machine', presented at the Third International Conference on Communications and Mobile Computing, Washington DC, USA, 18–20 April 2011, pp. 51–54.

[5]. K. Zheng, X. Qian, and N. An, 'Supervised Non-Linear Dimensionality Reduction Techniques for Classification in Intrusion Detection', presented at the International Conference on Artificial Intelligence and Computational Intelligence, Sanya, China, 23–24 October 2010, pp. 438–442.
[6] Townsend *et al*., 'k-NN Text Classification using an FPGA-Based Sparse Matrix Vector Multiplication Accelerator', presented at the IEEE International Conference on Electro Information Technology (EIT), Dekalb, IL, USA, 21–23 May 2015.
[7] J. Yang, X. Chen, X. Xiang, and J. Wan, 'HIDS-DT: An Effective Hybrid Intrusion Detection System Based on Decision Tree', presented at the International Conference on Communications and Mobile Computing, Shenzhen, China, 12–14 April 2010, pp. 70–75.
[8] S. Thaseen and C.A. Kumar, 'An Analysis of Supervised Tree Based Classifiers for Intrusion Detection System', in *Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering* (PRIME), Salem, India, 21–22 February 2013, pp. 294–299.
[9] N. Relan and D.R. Patil, 'Implementation of Network Intrusion Detection System using Variant of Decision Tree Algorithm', presented at the International Conference on Nascent Technologies in the Engineering Field (ICNTE), Navi Mumbai, India, 9–10 January 2015, pp. 1–5.
[10] G. Zhao, C. Zhang, and L. Zheng, 'Intrusion Detection using Deep Belief Network and Probabilistic Neural Network', presented at the IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), Guangzhou, China, 21–24 July 2017, pp. 639–642.
[11] N. Shone, T.N. Ngoc, V.D. Phai, and Q. Shi, 'A Deep Learning Approach to Network Intrusion Detection', in *IEEE Transactions On Emerging Topics In Computational Intelligence*, vol. 2, no. 1, February 2018, pp. 41–50.
[12] W. Zhong, N. Yu, and C. Ai, 'Applying Big Data Based Deep Learning System to Intrusion Detection', in *Big Data Mining and Analytics*, vol. 3, no. 3, September 2020, pp. 181–195.
[13] S. Ho et al., 'A Novel Intrusion Detection Model for Detecting Known and Innovative Cyberattacks using Convolutional Neural Network', in *IEEE Open Journal of the Computer Society*, vol. 2, January 2021, pp. 14–25.
[14] M. Lopez-Martin, B. Carro, and A. Sanchez-Esguevillas, 'Application of deep reinforcement learning to intrusion detection for supervised problems', in *Expert Systems with Applications*, vol. 141, Art. no. 112963, March 2020.
[15] S. M. Kasongo and Y. Sun, 'A Deep Long Short-Term Memory based classifier for Wireless Intrusion Detection System', in ICT Express Volume 6 No. 2, June 2020 pp. 98–103.
[16] C. Zhang, F. Ruan, L. Yin, X. Chen, L. Zhai, and F. Liu, 'A Deep Learning Approach for Network Intrusion Detection Based on NSL-KDD Dataset', presented at the IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID), 25–27 October 2019, pp. 41–45.
[17] T. Su, H. Sun, J. Zhu, S. Wang, and Y. Li, 'BAT: Deep Learning Methods on Network Intrusion Detection Using NSL-KDD Dataset', IEEE Access, vol. 8, February 2020, pp. 29575–29585.
[18] J. McHugh, 'Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory', in *ACM Transactions on Information and System Security*, vol. 3, no. 4, November 2000, pp. 262–294.
[19] C. Brown, A. Cowperthwaite, A. Hijazi, and A. Somayaji, 'Analysis of the 1999 DARPA/Lincoln Laboratory IDS Evaluation Data with NetADHICT', in *Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications*, Ottawa, ON, Canada, 8-10 July 2009.
[20] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, 'Toward developing a systematic approach to generate benchmark datasets for intrusion detection', *Computers & Security*, vol. 31, May 2012, pp. 357–374.
[21] S. T. Ikram and A. K. Cherukuri, 'Improving Accuracy of Intrusion Detection Model Using PCA and Optimized SVM', *Journal of Computing and Information Technology*, vol. 24, no. 2, June 2016, pp. 133–148.
[22] M. Ring et.al, 'A Survey of Network-based Intrusion Detection Data Sets', *Computers & Security*, vol. 86, September 2019, pp. 147–167.
[23] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, 'Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization', presented at the *International Conference on Information Systems Security and Privacy* (ICISSP), 2018, pp. 108–116.
[24] M. Ring, S. Wunderlich, D. Grüdl, D. Landes, and A. Hotho, 'Flow-based benchmark data sets for intrusion detection', in *European Conference on Cyber Warfare and Security* (ECCWS), 2017, pp. 361–369.
[25] G. Maciá-Fernández, J. Camacho, R. Magán-Carrión, P. García- Teodoro, and R. Therón, 'UGR'16: A New Dataset for the Evaluation of Cyclostationarity-Based Network IDSs', in *Computers & Security*, vol. 73, 2018, pp. 411–424.

[26] N. Moustafa and J. Slay, 'UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems', in *Military Communications and Information Systems Conference* (MilCIS), 2015, pp. 1–6.
[27] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, 'A Detailed Analysis of the KDD CUP 99 Data Set', in *Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications*, 2009, pp. 1–6.
[28] UCI KDD Archive Information and Computer Science, 'KDDCUP 1999 Data'. University of California, Irvine http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (accessed Oct. 19, 2021).
[29] Github, 'Coburg Intrusion Detection Data Sets', https://github.com/markusring/CIDDS (accessed Oct. 19, 2021).
[30] D. S. Henshel, G. M. Deckard, B. Lufkin, N. Buchler, B. Hoffman, P. Rajivan, and S. Collman, 'Predicting proficiency in cyber defense team exercises', in 2*016 IEEE Military Communications Conference (MILCOM)*, November 2016, pp. 776–781.
[31] D. Tovarňák, S. Špaček, and J. Vykopal, 'Traffic and log data captured during a cyber defense exercise', *Data in Brief*, vol. 31, August 2020.
[32] M. Pihelgas and M. Kont, 'Frankenstack: Real-time Cyberattack Detection and Feedback System for Technical Cyber Exercises', in *2021 IEEE CSR Workshop on Cyber Ranges and Security Training* (CRST), July 2021, pp. 396–402.
[33] D. Tovarňák, S. Špaček, and J. Vykopal, 'Dataset: Traffic and Log Data Captured During a Cyber Defense Exercise', *Zenodo*, April 2020, No. 105784.
[34] N. Känzig, R. Meier, L. Gambazzi, V. Lenders, and L. Vanbever, 'Machine Learning-based Detection of C&C Channels with a Focus on the Locked Shields Cyber Defense Exercise', in *11th International Conference on Cyber Conflict* (CyCon), 28–31 May, 2019, pp. 1–19.
[35] A. Kott, P. Théron, L. V. Mancini, E. Dushku, A. Panico, M. Drašar, B. LeBlanc, P. Losiewicz, A. Guarino, M. Pihelgas, and K. Rzadca, 'An introductory preview of Autonomous Intelligent Cyber-defense Agent reference architecture, release 2.0.', *The Journal of Defense Modeling and Simulation*, vol. 17(1), 2020, pp. 51–54.
[36] R. Meier, K. Heinäaro, V. Lenders, A. Lavrenovs, and L. Gambazzi, 'Towards an AI-powered Player in Cyber Defence Exercises', in *13th International Conference on Cyber Conflict* (CyCon), 2021, pp. 309–326.
[37] Defense Advanced Research Projects Agency. 'Cyber Grand Challenge (CGC)'. DARPA. https://www.darpa.mil/program/cyber-grand-challenge (accessed Nov. 5, 2021).

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

JARV1S: Phenotype Clone Search for Rapid Zero-Day Malware Triage and Functional Decomposition for Cyber Threat Intelligence

Christopher Molloy
Research Assistant
School of Computing
Queen's University
Kingston, ON, Canada
chris.molloy@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 Computing
Queen's University
Kingston, ON, Canada
ding@cs.queensu.ca

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

Abstract: Cyber threat intelligence (CTI) has become a critical component of the defense of organizations against the steady surge of cyber attacks. Malware is one of the most challenging problems for CTI, due to its prevalence, the massive number of variants, and the constantly changing threat actor behaviors. Currently, Malpedia has indexed 2,390 unique malware families, while the AVTEST Institute has recorded more than 166 million new unique malware samples in 2021. There exists a vast number of variants per malware family. Consequently, the signature-based representation of patterns and knowledge of legacy systems can no longer be generalized to detect future malware attacks. Machine learning-based solutions can match more variants. However, as a black-box approach, they lack the explainability and maintainability required by incident response teams.

There is thus an urgent need for a data-driven system that can abstract a future-proof,

human-friendly, systematic, actionable, and dependable knowledge representation from software artifacts from the past for more effective and insightful malware triage. In this paper, we present the first phenotype-based malware decomposition system for quick malware triage that is effective against malware variants. We define phenotypes as directly observable characteristics such as code fragments, constants, functions, and strings. Malware development rarely starts from scratch, and there are many reused components and code fragments. The target under investigation is decomposed into known phenotypes that are mapped to known malware families, malware behaviors, and Advanced Persistent Threat (APT) groups. The implemented system provides visualizable phenotypes through an interactive tree map, helping the cyber analysts to navigate through the decomposition results. We evaluated our system on 200,000 malware samples, 100,000 benign samples, and a malware family with over 27,284 variants. The results indicate our system is scalable, efficient, and effective against zero-day malware and new variants of known families.

Keywords: *malware analysis, malware triage, static analysis, binary clone search, information retrieval*

1. INTRODUCTION

Code reuse has been a common practice and strategy in software engineering, given the available vast open-source repositories [1] and the time and resources required to develop new code from scratch. The code that makes up the Advanced Persistent Threat (APT) groups that drive the development of malware is no exception to this practice of code reuse. Therefore, thanks to the readily available open-source components, code generation tools, and Malware-as-a-Service (MaaS) platforms, malware code, like legitimate code, is no longer created from scratch. Samples of malware that exhibit similar high-level behaviors are typically considered variants of the same malware family. Despite being created using code mutation and code obfuscation techniques to evade detection tools [2], malware variants in the same malware family share common code to achieve the same or similar sets of behavior or action sequences. In recent years, malware variants have exponentially increased in both volume and threat potential. Emotet, a popular ransomware tool with more than 70,000 variants, has caused upward of USD 1 million per incident for governments and private institutions.[1, 2] According to the AV-TEST Institute, 21.8 million new malware variants that target machines running Microsoft operating systems were recorded in 2021.[3] Due to the ever-increasing number of malware variants, malware samples are categorized into families based on their code, runtime behavior, and functionality.

1 https://www.malwarebytes.com/emotet
2 https://bazaar.abuse.ch/browse/tag/Emotet/
3 https://www.av-test.org/en/statistics/malware/

Matching unknown variants to the correct malware families and recognizing novel malware from the same unknown families is a critical process for a successful malware triage in cyber threat intelligence (CTI). Typical methods of malware analysis include dynamic analysis, signature-based pattern matching, and machine learning-based detection methods. Dynamic analysis is done by observing the behavior of a malware sample in an isolated virtual environment [3] by executing the malware sample itself. The behavior of the malware is recorded and compared with that of other malware variants in the behavior database. However, modern malware samples have been equipped with anti-sandboxing techniques that can recognize when they are executed in a virtual environment and avoid executing their malicious payload [4]. Additionally, dynamic analysis is a time-consuming process that can take up to two minutes per malware sample [3], which fails the time-sensitivity requirement of a CTI pipeline.

Typical anti-virus software, such as McAfee and Microsoft Defender, uses a signature-based approach for malware classification. Signature-based pattern matching methods extract a small string of identifiable bytes from unknown malware and compare it to known malware signatures [5]. Although signature-based methods can identify malware from known families with a low error rate, malware authors can implement obfuscation and packing techniques to impede the signature extraction process [5]. To address this problem, machine learning solutions have been developed for malware clone searching [6]–[9]. Although these methods have been shown to work with a high number of variants, they do not provide any explainability for their classification results, and the models persistently need to be re-trained and tuned for new malware families that use new code obfuscation and packing techniques. APTs behind the malware attacks are a continual and evolving threat to organizations and governments. To ensure a strong defense against these constantly changing threats, we must keep moving against the adversaries in defense development.

We propose a different approach – Phenotype Clone Search for Functional Decomposition. We use phenotypes as the basic elements of observable characteristics extracted from a given malware sample. Examples of phenotypes are code fragments, strings, and constants. They can all be directly observed without sandboxing or emulation. Since a phenotype is defined as a small observable piece of information, humans can easily interpret its meaning. Malware variants in a given malware family share a similar codebase, and therefore, they will share many phenotypes. The organization and presentation of phenotypes may change from sample to sample, but the basic elements of a malware executable cannot easily be hidden. In our proposed approach, we define the basic unit of malware functionality as an assembly function extracted from a malware sample. As a unit of functionality, an assembly function contains a set of observable phenotypes. By matching phenotypes, we can search for cloned units of malware functionality (in this case, the cloned assembly functions)

and find shared functionality in different malware samples. Given a target piece of malware under investigation and a repository of known malware, the objective is to decompose the target into known units of malware functionality from the repository. Figure 1 shows an example. The target under analysis is a sample of the Dtrack malware family. Dtrack is a Remote Administration Tool (RAT) malware that has been found in a nuclear power plant.[4] Our approach decomposes the functions of the target malware into known functions from the repository. We found that about 11,000 of the functions come from common code that can be found in benign software. Of the rest of the functions, 563 can be found in functions from samples of the same family. Some common malware code that is shared with other families was also found. For zero-day malware samples that do not belong to any known family, we can also study the compositional relationship by searching for them in the repository. Our contributions can be summarized as follows:

- We proposed a novel approach to malware functional decomposition to complement existing dynamic, signature-based, and machine learning-based malware triage solutions.
- We designed a new distributed clone search-based analytic framework that enables an efficient, accurate, and robust malware decomposition implementation.
- The proposed solution can decompose a given piece of target malware into known functionality with interactive visualization, scalable to malware samples of exceptionally large size.

The rest of this paper is organized as follows. Section 2 describes the overall design of the JARV1S system as the host to the proposed malware decomposition method. Section 3 elaborates the clone search method and how we use it to decompose malware functionality. Section 4 demonstrates the effectiveness of our system. Section 5 discusses the related research in this domain. Finally, Section 6 provides the conclusion.

2. SYSTEM DESIGN

The proposed JARV1S system itself is an extendable open design platform that aims to host scalable and efficient malware analysis algorithms of diverse types and accommodate their special needs. There are several design considerations. The first one concerns storage systems and job execution workers. The scale of the data to be processed means that the limited resources of a single server are unlikely to offer enough capacity. Additionally, a single server suffers from the problem of a single point of failure. We could leverage existing distributed data storage options.

4 Dtrack: In-depth analysis of APT on a nuclear power plant.

However, typically the distributed storage nodes and the nodes used for job execution are in separate clusters of nodes. This implies that one would need to maintain at least three separate clusters: storage nodes, execution nodes, and messaging queue nodes. A typical way to improve cluster performance and data throughput is via data locality: a given job is assigned to a worker which is co-located with the data in the same node. Separating the computation nodes and storage nodes would remove the advantage of data locality. Therefore, we propose implementing our own distributed storage and job execution distributed cluster. Since we know our data models and the attributes of the information extracted from a malware artifact, and they are unlikely to change significantly in the future, the storage system can be further simplified to better balance the retrieval speed versus the compression ratio. Vision, the storage and job execution cluster of JARV1S, contains three components:

- *Load balancer and remote procedure call (RPC) proxy*. This is the main interface between the Vision cluster and the SocketIO-based web frontend. The load balancer keeps track of the storage and execution load of each node, assigns new project storage requests, and manages replicas if configured.

- *Vision node*. A Vision node contains a list of projects stored in folders similar to a Git[5] repository (see Figure 2). Each project contains a list of user-uploaded artifacts, extracted information, and analysis reports. Associated with a given project folder is a Python-based API (application programming interface) for asset uploading, unpacking, disassembling, and running analyses. The API is served through a remote proxy and load balancer. As a project is physically stored in a folder, it must fit into a single node. A Vision node also contains a list of Huey workers.[6] These executors will receive asynchronized job requests and directly operate on the projects stored in the same node. Job requests will be assigned to executors that are on the same node as the target project. However, if all the executors on the given project's node are busy, a job request will be assigned to a worker on a different node, and the worker will access the project by calling remote procedures via the proxy.

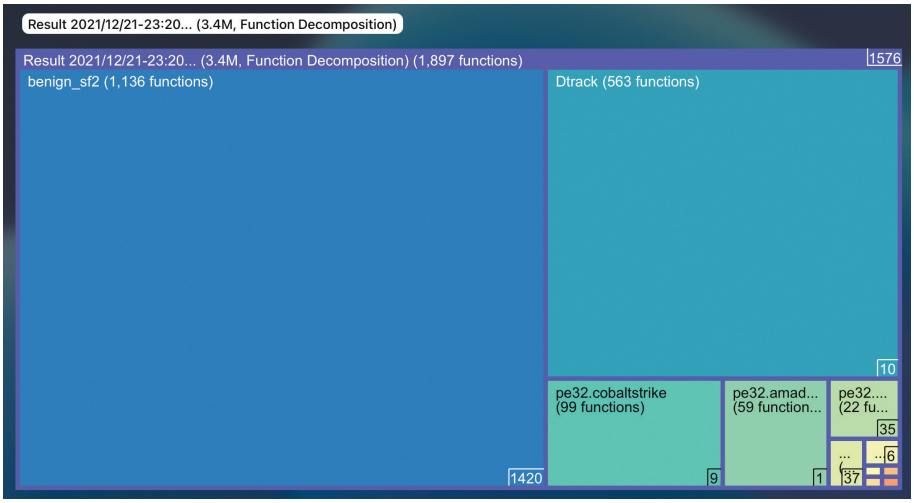
- *GPU-enabled Vision node*. Vision nodes in this category have one or more GPU devices available for machine learning-based analytic tasks. However, GPU-related tasks will have priority to take over the GPU-enabled job workers.

The second consideration is the user interface adaptability and responsiveness. In our past solutions, we generated a single report in the format of a JSON document

5 Git source code version control framework.
6 Huey: a lightweight job executor framework.
7 VirusTotal: a free virus, malware, and URL online scanning service.
8 Hybrid Analysis: a free malware analysis service.

and rendered it directly through a website. This is also the typical practice used in existing CTI online platforms for malware analysis and triage, such as VirusTotal[7] by Google and Hybrid Analysis[8] by CrowdStrike. However, as internet connection speeds have significantly improved, attackers have started to increase the size of their malware artifacts. For example, we have seen malware samples that exceed 200 MB, and an analysis report for such a sample could exceed 1 GB. Rendering a 1 GB JSON document with diagrams and visualizations on the client-side is infeasible. The existing frameworks simply skip files of a certain size. For instance, Hybrid Analysis has a maximum file size of 100 MB. To address this issue, we represent each analysis report as an SQLite[9] database stored in a single file. Therefore, we can load and render specific information on the web interface as needed based on queries to the SQLite database. We use SocketIO,[10] a room-based event-driven framework, as the channel for client-frontend and client-Vision communication to transfer information about the project and the analytic report.

FIGURE 1: AN EXAMPLE OF MALWARE FUNCTIONAL DECOMPOSITION. GIVEN A MALWARE SAMPLE UNDER ANALYSIS (TRACK FAMILY), OUR SYSTEM EFFICIENTLY BREAKS DOWN THE SAMPLE INTO GROUPS OF KNOWN UNITS OF FUNCTIONALITY FROM THE REPOSITORY



The third consideration is plug-in integration. Direct integration of reverse engineering tools such as IDA Pro[11] and Ghidra[12] is highly convenient for malware analysts, as they can obtain the analytic report on the fly while they are investigating certain malware campaigns. However, these disassemblers come with different plug-in requirements, especially regarding the integration of user element components. These requirements and related APIs also frequently change. Instead, we propose connecting

9 SQLite Database: a small, fast, and self-contained database.
10 Socket.IO: Bidirectional and low-latency communication.
11 IDA Pro: an interactive disassembler.
12 Ghidra: A software reverse engineering (SRE) suite of tools.

the disassembler plug-in to the web frontend as a SocketIO-based remote procedure call following the same project API interface we used in the Vision node. The user will directly use the web user interface (UI) of the frontend service to conduct analysis and access results (see Figure 3).

FIGURE 2: JARV1S SYSTEM COMPONENTS AND THE FLOW DIAGRAM FOR THE MALWARE EXTRACTION AND INDEXING PROCESS

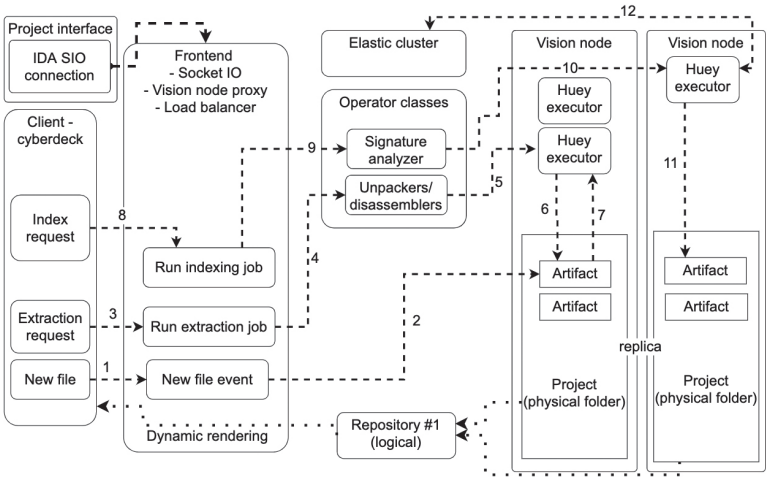
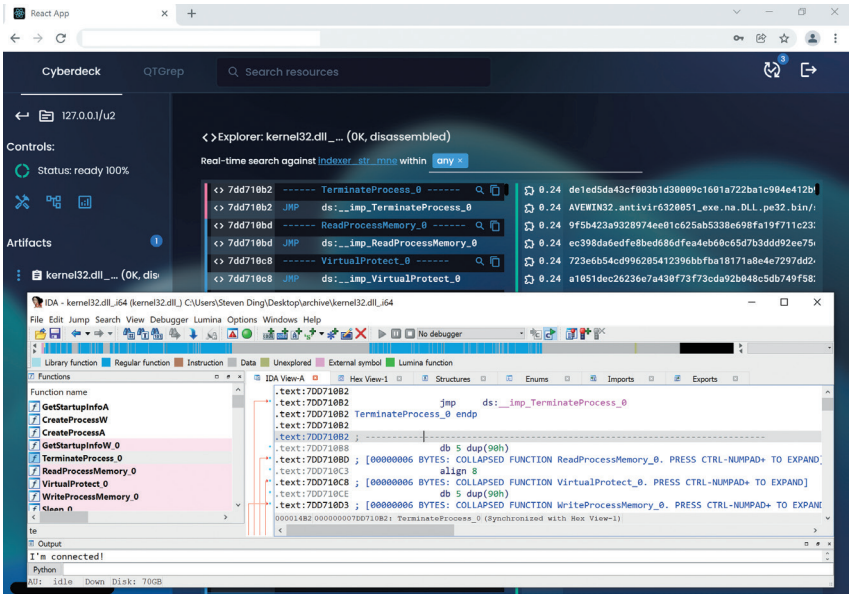


FIGURE 3: IDA PRO PLUG-IN CONNECTED TO THE FRONTEND SERVICE AND SHOWN UP ON THE WEB INTERFACE. THE WHITE WINDOW AT THE FRONT IS IDA PRO AND THE BROWSER WINDOW AT THE BACK IS THE JARV1S WEB UI. BOTH SHOW THE SAME INFORMATION SYNCED BY SOCKETIO



The project-based analytic and data retrieval interface is, therefore, backed by three different implementations:

- *A Python-based implementation that directly accesses the stored data in a folder*. This implementation is used for the data-locality mode, where the job worker is assigned to the same Vision node as the stored project folder.

- *An RPC-based object proxy implementation*. The executed code is the same as for the Python-based implementation, except the APIs are called from a different Vision node than the one in which the project folder is stored. The API is triggered when the job worker is located in a different physical node to the project folder.

- *A SocketIO-based RPC object proxy implementation*. This is used as the plug-in-frontend and plug-in-Vision communication channel. It is separate from the RPC-based object proxy implementation, as the communication channel to the client and the plug-in is untrusted and needs authentication and certain protections. Thus, the frontend can directly call a project API to retrieve information which is then rendered on a webpage. For example, if a list of function names of a certain range is requested, the list can be rendered in the user's browser. The Vision node can also connect to the same SocketIO room, identified by the user's unique id, to access any of the SocketIO-connected plug-in clients.

3. MALWARE DECOMPOSITION

The clone search-based malware decomposition method consists of two procedures: indexing and analyzing. Figure 2 shows the steps involved for a user to index an artifact or a project which contains a list of artifacts, specifically:

1. The user submits one or more malware samples through the web browser into an existing or newly created project. The new file event is triggered within the SocketIO room specifically for that user. The samples are transferred to the frontend SocketIO server in small chunks.
2. Upon receiving the complete file samples, the frontend server resolves to the Vision cluster load balancer and RPC proxy. Files are then transferred to the assigned Vision node, compressed, and stored in the corresponding project folder.
3. The user submits an extraction request through the client interface. The SocketIO server triggers the event to execute the extraction.

4. The triggered event is mapped into the operator classes for all the applicable extraction steps needed for the target resources. These operator classes include extraction, unpacking, disassembling, and optionally decompiling.
5. The mapped operator classes and the specific job configuration will then be submitted into the Vision cluster through a load balancer that considers both data locality and individual workloads.
6. The assigned job executor starts communicating with the project folder via the local project API, a SocketIO-backed project API, or an RPC-backed project API.
7. The job executor reads the necessary information about the artifacts and writes the extracted information back to the project folder. The disassembly code, strings, constants, and other common attributes of a malware sample are saved to an SQLite single-file database with a customized compression dictionary. It should be noted that each executor will be assigned a single artifact to extract, rather than all the artifacts to increase parallelism.
8. Once the extraction process is complete, the user can submit an indexing job through the web user interface on the client.
9. Like the previous user actions, a new event specifically for the indexing job is fired from the SocketIO interface. The event is then mapped into the corresponding analyzers under the operator classes. These analyzers define any additional extraction and analysis steps needed before making an artifact searchable.
10. The analyzer class and the configurations will be submitted to the Vision node, and a job executor will be allocated for the submitted task. It should be noted that to provide efficient access and search, certain frequently accessed projects will be replicated across different nodes through synchronized project API transactions, to increase data parallelism. Replications are designed to be automatically provisioned. Once a project is accessed less frequently, its replica(s) will be removed.
11. During the decomposition process, the analyzer will extract the observable characteristics of code, data, and strings from the disassembly code. It will then submit the extracted phenotypes as basic indexable elements into a separate Elasticsearch cluster.
12. Finally, the job executor will update the project status and indicate the job has been completed.

To conduct a decomposition analysis of a specific project, a similar list of steps is followed, as shown in Figure 4. The user first submits an analysis request through the user interface (Step 8), targeting a specific project. For new artifacts that do not yet exist in the system, the user follows Steps 1–7 from Figure 2 to create a new project and add new artifacts. It should be noted that the target does not have to be indexed.

After Step 8 (Figure 4), the job analysis event is fired, mapped into the corresponding analyzer classes, and then submitted to the Vision cluster, where a job executor is requested (Steps 9–10). Once assigned the job, the executor will extract phenotypes and run the decomposition algorithm. The analytic result is saved to a single-file SQLite database (Steps 11–13).

The proposed functional decomposition method is a three-step process: (1) given a target malware sample, we identify the list of assembly functions. (2) For each assembly function, we search for its clones in a collection of benign software binaries to remove some commonly reused library code. (3) For the assembly functions that do not match any existing benign library functions, we conduct another clone search against the malware in the repository. Overall, the basic assembly code clone search is combined and repeated at the function level. To search for clones of given assembly functions, we opted for a signature-based approach. As mentioned earlier, we extract phenotypes, that is, observable characteristics from the assembly code, to match functional level clones.

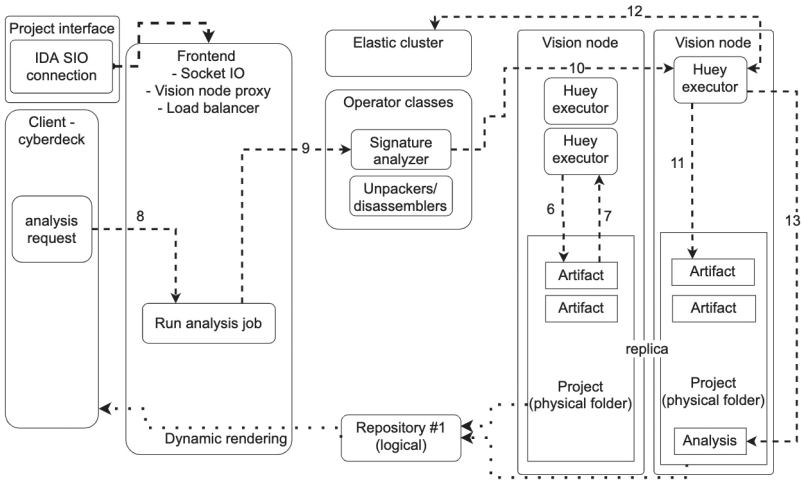
- Assembly code operation n-grams: Given a piece of assembly code, we extract the list of assembly operations, such as *mov, add, push,* and *pop*. Then, we extract sequential *n* operations as an *n*-gram single phenotype. For example, (*mov, add, add*) is a 3-gram phenotype. We use both 2-grams and 3-gram phenotypes. These n-gram sizes were chosen based on results by Khoo *et al*. [14] and our empirical evaluation.

- Referred string constants and stack strings: Given a piece of assembly code, we identify a list of referred string constants as phenotypes. We investigate the data reference operations and operands in the assembly code, such as data loading from a specific segment. Given the located data reference address, we scan and verify if there is a valid UTF-8 or ASCII string present by checking if the encoding is valid. For malware samples, static strings are often obfuscated as stack strings, where each character of the string is stored separately and combined at run time. We identify short strings from a single basic block and merge them into a single string value.

- Numeric literals: Numeric literals are useful to match different implementations or variants of the same compression or encryption algorithm. Figure 5 shows an example of the adler32 function, where the 0xfff1 constant is critical for its control flow. We extract all the referred constant numbers from the given piece of assembly code. Numeric literals can be found in different forms in assembly code. The same number can be encoded in a different format or be of a distinct size. We normalize the numbers according

to their type. Integers in base 2, 8, 10, or 16 are converted into integer values in base 16 in the form of a hexadecimal string. We also include its original form in case it is part of the disassembler analysis comments, where proper reference resolution and value normalization have already been completed. For floating-point values, it is difficult to match two identical numbers if they are encoded with a different precision schema. Instead, we convert each floating-point number into two hex strings, one for single-precision encoding and one for double-precision encoding, as the phenotypes will match to any format.

FIGURE 4: JARV1S SYSTEM FLOW DIAGRAM FOR THE MALWARE DECOMPOSITION PROCESS



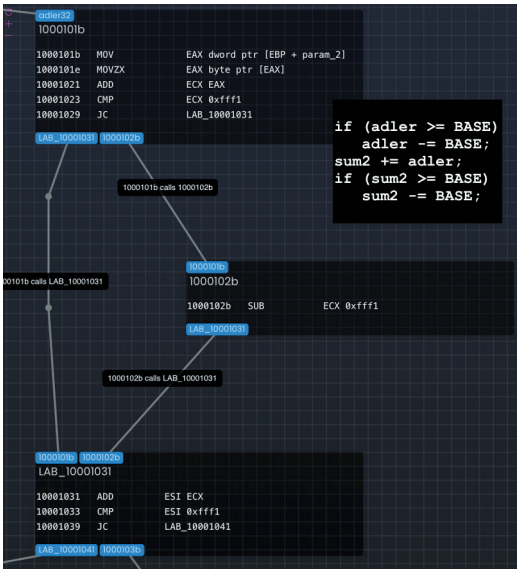
Given a target assembly function, we extract a set of phenotypes and match it against the repository by combing the Okapi bm25 function and the TF-IDF function [10]. The Okapi bm25 scoring function is popular for text data retrieval, where each piece of text data to be retrieved is represented as a set of keywords. Given a query Q and a candidate document D, the similarity score can be formulated as:

$$\mathrm{bm25}(D,Q) = \sum_{i=1}^{n} \mathrm{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot(1-b+b\cdot|D|/\mathrm{avgdl})} \tag{1}$$

This function is a cumulative sum of independent scores evaluated for each of the n query terms q_i. In this context, it is one of the extracted phenotypes, such as a single operation 2-gram. For each phenotype q_i, IDF(q_i) measures the inverse of the total number of assembly functions that have q_i. A common phenotype q_i, which is presented in many assembly functions, will thus contribute less to the total score.

$f(q_i, D)$ measures the frequency of phenotype q_i in the candidate function D. The more frequently the phenotype appears in the candidate, the more representative it is for that candidate. The frequency is normalized by the relative length of the given candidate function compared to the average length of the assembly functions in the repository. k_1 and b are two hyperparameters controlling the degree of normalization. *avgdl* is the averaged document length in the database. A longer candidate function D tends to have higher frequency counts for the extracted phenotype. This is especially important to match for certain compiler optimizations, where loops are unrolled. k_1 and b control the degree of normalization strength at the query time.

FIGURE 5: VISUALIZING THE CONTROL FLOW GRAPH OF A FUNCTION IN JARV1S. THIS EXAMPLE SHOWS THAT 0XFFF1 IS A CRITICAL NUMERIC LITERAL FOR THE ADLER32 FUNCTION. THE WHITE TEXT IS THE ORIGINAL SOURCE CODE



Given a query Q, the scores for all the candidates D are relative scores. This means that the candidate's score can only be compared within the result of the same query and its range is unbounded. To bound the score into a human-friendly similarity score of range 0 to 1, we further normalize the score and divide a candidate's score by a bm25 score of matching the query to itself.

$$\mathrm{bm25n}(D, Q) = \frac{\mathrm{bm25}(D, Q)}{\mathrm{bm25}(Q, Q)} \tag{2}$$

This way, the candidate score is bound to 1, and it is comparable across different

queries. bm25 is the default scoring function for Elasticsearch, and it is ideal for situations where the queries are mostly shorter than the indexed documents. This is quite a common situation for text data retrieval in search engines. The queries mostly consist of only a few words, and in contrast, the retrieved documents are much longer, such as a news article. In this case, for phenotype matching in an assembly clone search, bm25 is a useful scoring function against compiler optimization techniques such as loop unrolling and inline function calls, where the query is indeed much shorter than the correct candidate to be matched.

However, there are other situations where the bm25 function does not fit well. If the query and the candidates are of comparable length, bm25 may not be ideal, due to the normalization term for D. In this scenario, for text data retrieval, the TF-IDF scoring function is more applicable. It can be formulated as:

$$\text{tfidf}(D, Q) = \sum_{i=1}^{n} sqrt(f(q_i, D)) \cdot log(\frac{DF + 1}{\text{DF}(q_i) + 1}) \cdot \frac{1}{sqrt(|D|)} \tag{3}$$

Where q_i D, and $f(q_i, D)$ are the same as above. DF and $DF(q_i)$ are the document frequency and the frequency of sample q_i in the document respectively. Like the bm25 score function, the TF-IDF score function is also an unbounded relative score. Following the same approach, we normalize the TF-IDF score by dividing it with a self-match TF-IDF score:

$$\text{tfidfn}(D, Q) = \frac{\text{tfidf}(D, Q)}{\text{tfidf}(Q, Q)} \tag{4}$$

Finally, we combine both scoring functions into a unified similarity score for candidate ranking by considering the discriminative power in the retrieved candidates:

$$d(Q, s) = D_{KL}(\max_{k}\{s(d_i : i = 1..|\mathbb{D}|)\} \;||\; \underbrace{\{1, 0, 0, ..., 0\}}_{k \text{ elements}}) \tag{5}$$

$$D_{KL}(a \;||\; b) = \sum_{i} a_i \cdot log2(a_i / b_i) \tag{6}$$

Given a query Q, we measure the discriminative power of $s \in$ {bm25n, tfidfn} by comparing the KL-divergence between the top-k retrieved candidate scores, where k is the user-defined parameter controlling the maximum number of returned results, and a one-hot vector of k elements where only the first element is 1. An ideal scoring function in an ideal situation should be able to provide good discriminative power over the list of candidates while only keeping the first candidate a perfect matching score. Then, we select the scoring function that yields the lowest value of d:

$$\text{similarity}(D,Q) = \begin{cases} \text{bm25n}(D,Q), & \text{if } d(Q,\text{tfidfn}) > d(Q,\text{bm25n}) \\ \text{tfidfn}(D,Q), & \text{otherwise} \end{cases} \tag{7}$$

As discussed earlier, the decomposition process consists of three main steps. Algorithm 1 provides the details of each individual step. Line 1 corresponds to Step 1 for assembly function extraction and retrieval. Lines 2–8 correspond to Step 2 for the clone search against a list of benign artifact projects. The goal is to first remove any known functions commonly existing in the benign executables. Lines 10–18 correspond to Step 3, where we search for unknown functions in the malware projects.

Algorithm 1 Functional decomposition of a sample s by clone search

```
1: 𝕼 ← extract(s)   ▷ Retrieve assembly functions from the artifact s after disassembling
2: unknown = {}
3: for Q ∈ 𝕼 do                                   ▷ Loop through each extracted function Q.
4:     m ← max_1{similarity(Q, D) : D ∈ 𝔹}          ▷ Clone search against benign set.
5:     if m != 1 then          ▷ If m equals 1, we found a known benign function.
6:         unknown = {m, ...unknown}                  ▷ Store the unknown function.
7:     end if
8: end for
9: decomposition ← {}                              ▷ The decomposition result mapping.
10: for u ∈ unknown do                        ▷ Loop through each unknown function u.
11:     m ← max_k{similarity(u, D) : D ∈ 𝕄}      ▷ Clone search against the malicious set.
12:     for r ∈ m do                                  ▷ Loop through each candidate.
13:         if r.score == m_0.score then      ▷ Consider entries having the same top score.
14:             p ← project(r)                        ▷ Retrieve candidate's project ID.
15:             decomposition[p] = (u, r)  ▷ Store the entry result entry under given project.
16:         end if
17:     end for
18: end for
19: return decomposition
```

4. EXPERIMENTS

To test our design, we created a repository of 200,000 malware samples along with 100,000 benign samples. Malware samples are grouped based on their family, and the samples of the same family are stored in the same Vision project. All benign samples are stored in the same project. It takes around 43 hours in total to unpack, disassemble, and index all the samples. Afterwards, we use a separate set of malware families to simulate zero-day malware families and samples that were not indexed into the system. The malware repository contains 394 families, and the repository of benign binaries contains 4,098 categories. Table I lists some of them. To simulate the zero-day malware families, we use a list of families that are not indexed in the repository.

For each sample in the zero-day malware family set, we search and attempt to match it against all the samples, including the benign repository, the malware repository, and the zero-day set, using malware decomposition analysis. Given an analytic result, we

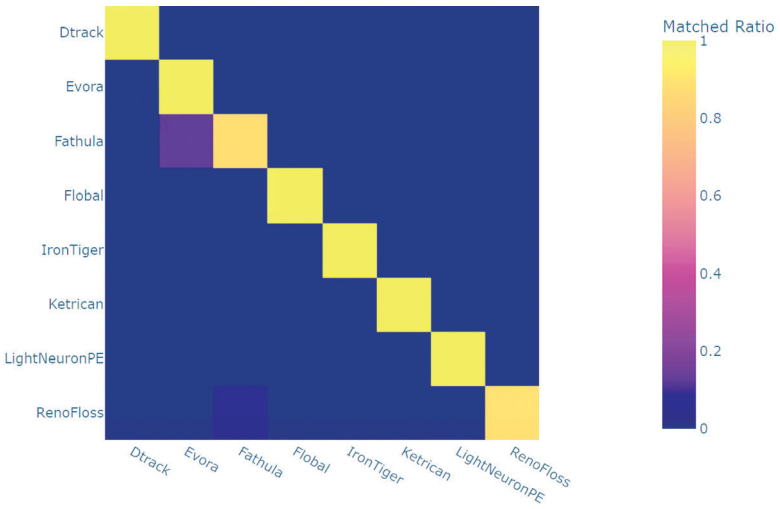
remove all the detected benign assembly functions and check to see to what family the majority of the rest of the assembly functions belong. Figure 6 shows an example. In total, there are 356 functions after removing the 35 benign functions. There are 156 functions that matched the Evora family and 145 functions that matched the Fathula family (no overlapping functions). Therefore, the top-matched family for this sample is Evora. Given all the samples of a malware family, we estimate the ratio of samples matched for each of the zero-day families. Figure 7 shows the matching ratio across different families and across the zero-day set. It shows that the majority of the malware families were correctly matched (100% ratio). There were two false negatives (15%) for the Fathula family and one false negative (5%) for the RenoFloss family.

TABLE I: SOME SAMPLE MALWARE FAMILIES AND BENIGN SOFTWARE CATEGORIES USED IN THE EXPERIMENT

Dataset	Malware Families
Malware Repository	Agent Tesla, Formbook, GuLoader, NanoCore, RemcosRAT, MASS Logger, NjRAT, Dridex, Quakbot, Gozi, AveMariaRAT, Snake Keylogger, IcedID, AsyncRAT, RedLine Stealer, HawkEye Keylogger, Cobalt Strike, RaccoonStealer, NetWire, ModiLoader...
Benign Repository	Winamp, Mozilla Firefox, Maxthon, SeaMonkey, Windows Live Messager, Pidgin, ESET Nod32, Defraggler, Avant Browser, TortoiseSVN, Picasa2, Google Talk, VirtualBox, Thunderbird, Silverlight, PowerDVD, FastStone Image Viewer, PowerISO, Flash Player, 7z, MySQL, Blender, XBMC Media Center, Win7codecs, Tera Term Pro...
Zero-day Set	Dtrack, Evora, Fathula, Flobal, IronTiger, Ketrican, LightNeuronPE, RenoFloss

FIGURE 6: TREE MAP VISUALIZATION OF THE ANALYTIC SUMMARY FOR A SAMPLE FROM THE EVORA FAMILY

FIGURE 7: SIMILARITY MATRIX REGARDING THE MATCH RATIO IN THE ZERO-DAY FAMILY TESTING SET



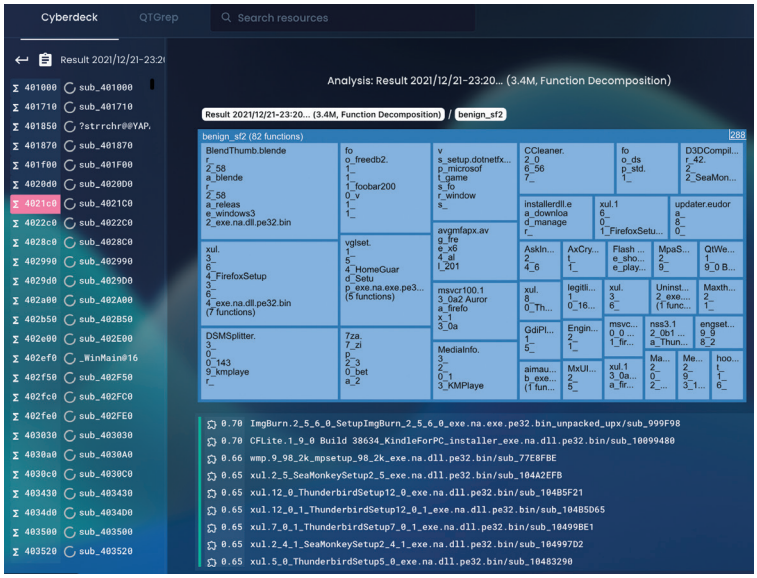
Besides the tree map view shown in Figures 1 and 6, our UI also allows the user to further explore the detected cloned functions (Figure 8). The user can browse the list of assembly functions (left panel). After clicking on one of the entries, the user can see the list of detected function clones on the lower right panel. The user can also compare and see the differences between the selected assembly function and the candidate function by clicking on one of the entries. The differences are visualized using either a typical text comparison method or that combined with a graph-based comparison method. Additionally, the user can click on any of the rectangles shown on the tree map to further break down the detected clones of a specific family to see how the clones are distributed across samples in a specific family. The upper-right panel in Figure 8 shows an example of clicking into the benign group and seeing how the functions of different benign samples can be matched.

5. RELATED WORKS

Malware similarity analysis is the area of research methods for matching malware variants. Similarity analysis research has shown multiple different techniques for successful variant matching, such as feature hashing, neural network-based similarity analysis, and clone detection. One method proposed by Jang *et al*. used hashed feature vectors and co-clustering techniques for malware detection and family classification [11]. In a real-world environment, there is no way to predict which features will be

significant in the future. This is why JARV1S bases its analysis on assembly code. Neural networks have also been applied to malware family categorization. Works proposed by Hsiao *et al*., Zhu *et al*., Stokes *et al*. and Khandhar all used Siamese-based neural network architectures for malware family classification [6]–[9]. Unlike JARV1S, these methods either fail to show any results against zero-day malware or show results with decreasing accuracy.

FIGURE 8: THE USER INTERFACE FOR RENDERING OUR DECOMPOSITION ANALYSIS RESULT



Clone searching is a method of searching for duplicate code fragments in multiple software repositories [12]. Clone searching for benign software has been researched due to the frequent practice of copying and pasting code with minor changes. Recently, clone searching has expanded to analyze assembly code [13]–[17]. Khoo *et al*. identified code clones using control flow sub-graphs and data constants [14]. Extracted semantic signatures were compared for clone detection by Hu *et al*. by emulating binary function execution [15]. Farhadi *et al*. proposed BinClone, a tool for malware code fragment clone detection using a token-level approach [16]. Asm2Vec, a framework to learn the lexical semantic relationships of assembly functions based on assembly code, was proposed by Ding *et al*. [13]. In addition, Ding *et al*. proposed the static framework Kam1n0 for assembly code-based clone search [17]. Unlike these systems, not only does JARV1S find the similarities between binaries but also compares the binaries in relation to their behavior and their connection to APTs. Furthermore, JARV1S provides a human-friendly environment for result analysis.

Xue *et al*. and Lee *et al*. have also conducted clone searches on assembly code, but the systems they proposed were designed specifically to find compromised code in a repository [18], [19].

6. CONCLUSION

In this paper, we proposed a new malware triage system. Given a piece of malware, it can trace the origin of the extracted assembly functions from known binary categories and samples. We proposed the concept of phenotype assembly clone search, where one can match different assembly functions based on their observable characteristics. The proposed system is designed with scalability, efficiency, and adaptability considerations. For real-world application, JARV1S has been designed for seamless integration into an existing APT processing pipeline. Our experiments show that the system is efficient and accurate in analyzing new malware samples from both known and unknown malware families.

REFERENCES

[1] A. Mockus, "Large-scale code reuse in open-source software," in *Proceedings of the First International Workshop on Emerging Trends in FLOSS Research and Development* (FLOSS'07: ICSE Workshops 2007), 2007, pp. 7–7.

[2] Y. Ye, T. Li, D. A. Adjeroh, and S. S. Iyengar, "A survey on malware detection using data mining techniques," *ACM Comput. Surv*., vol. 50, no. 3, pp. 41:1–41:40, 2017.

[3] M. Apel, C. Bockermann, and M. Meier, "Measuring similarity of malware behavior," in *Proceedings of the 34th Annual IEEE Conference on Local Computer Networks* (LCN 2009), Zurich, Switzerland, Oct. 20–23, 2009, pp. 891–898.

[4] B. Lau and V. Svajcer, "Measuring virtual machine detection in malware using DSD tracer," *J. Comput. Virol*., vol. 6, no. 3, pp. 181–195, 2010.

[5] Y. Ye, T. Li, S. Zhu, W. Zhuang, E. Tas, U. Gupta, and M. Abdulhayoglu, "Combining file content and file relations for cloud-based malware detection," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, C. Apté, J. Ghosh, and P. Smyth, Eds. San Diego, CA, USA, Aug. 21–24, 2011, pp. 222–230.

[6] J. Zhu, J. Jang-Jaccard, and P. A. Watters, "Multi-loss Siamese neural network with batch normalization layer for malware detection," *IEEE Access*, vol. 8, pp. 171542–171550, 2020.

[7] J. W. Stokes, C. Seifert, J. Li, and N. Hejazi, "Detection of prevalent malware families with deep learning," in *Proceedings of the 2019 IEEE Military Communications Conference* (MILCOM 2019), Norfolk, VA, USA, Nov. 12–14, 2019, pp. 1–8.

[8] S. Hsiao, D. Kao, Z. Liu, and R. Tso, "Malware image classification using one-shot learning with Siamese networks," in *Proceedings of Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019*, Budapest, Hungary, Sep. 4–6, 2019, in Procedia Computer Science, I. J. Rudas, J. Csirik, C. Toro, J. Botzheim, R. J. Howlett, and L. C. Jain, Eds., vol. 159, pp. 1863–1871.

[9] S. Khandhar, "A few-shot malware classification approach for unknown family recognition using malware feature visualization," M.S. thesis, Delft Univ. of Tech., Delft, The Netherlands, 2021.

[10] I. C. Mogotsi, "Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze: Introduction to information retrieval," *Information Retrieval*, vol. 13, pp. 192–195, 2010.

[11] J. Jang, D. Brumley, and S. Venkataraman, "Bitshred: feature hashing malware for scalable triage and semantic analysis," in *Proceedings of the 18th ACM Conference on Computer and Communications Security* (CCS 2011), Y. Chen, G. Danezis, and V. Shmatikov, Eds. Chicago, IL, USA, Oct. 17–21 2011, pp. 309–320.

[12] H. Zhang and K. Sakurai, "A survey of software clone detection from security perspective," *IEEE Access*, vol. 9, pp. 48157–48173, 2021.
[13] 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 *Proceedings of the 2019 IEEE Symposium on Security and Privacy* (SP 2019), San Francisco, CA, USA, May 19–23, 2019, pp. 472–489.
[14] W. M. Khoo, A. Mycroft, and R. J. Anderson, "Rendezvous: a search engine for binary code," in *Proceedings of the 10th Working Conference on Mining Software Repositories* (MSR '13), T. Zimmermann, M. D. Penta, and S. Kim, Eds. San Francisco, CA, USA, May 18–19, 2013, pp. 329–338.
[15] Y. Hu, Y. Zhang, J. Li, and D. Gu, "Binary code clone detection across architectures and compiling configurations," in *Proceedings of the 25th International Conference on Program Comprehension* (ICPC 2017), G. Scanniello, D. Lo, and A. Serebrenik, Eds. Buenos Aires, Argentina, May 22–23, 2017, pp. 88–98.
[16] M. R. Farhadi, B. C. M. Fung, P. Charland, and M. Debbabi, "Binclone: Detecting code clones in malware," in *Proceedings of the Eighth International Conference on Software Security and Reliability* (SERE 2014), San Francisco, CA, USA, Jun. 30 – Jul. 2, 2014, pp. 78–87.
[17] S. H. H. Ding, B. C. M. Fung, and P. Charland. "Kam1n0: MapReduce-based assembly clone search for reverse engineering," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (SIGKDD), San Francisco, CA, USA, August 2016, pp. 461–470.
[18] H. Xue, G. Venkataramani, and T. Lan, "Clone-slicer: Detecting domain specific binary code clones through program slicing," in *Proceedings of the 2018 Workshop on Forming an Ecosystem Around Software Transformation*, 2018, pp. 27–33.
[19] Y. J. Lee, S.-H. Choi, C. Kim, S.-H. Lim, and K.-W. Park, "Learning binary code with deep learning to detect software weakness," in *Proceedings of the 9th International Conference on Internet* (ICONI), 2017.

2022 14th International Conference on Cyber Conflict
Keep Moving
T. Jančárková, G. Visky, I. Winther (Eds.)
2022 © NATO CCDCOE Publications, Tallinn

Emergence of 5G Networks and Implications for Cyber Conflict

Keir Giles
Conflict Studies Research Centre
Northamptonshire, United Kingdom
keir.giles@conflictstudies.org.uk

Kim Hartmann
Conflict Studies Research Centre
Northamptonshire, United Kingdom
kim.hartmann@conflictstudies.org.uk

Abstract: The internet of things (IoT), autonomous driving, or Industry 4.0 – regardless of the application scenario envisioned, next-decade technologies reliant on connectivity will be based on 5G infrastructure and become increasingly dependent on virtualizations to provide adequate and adaptable network services. Virtualized network functions (VNFs) are used to provide services through software that replaces dedicated network devices. This shift from physical devices to software functions allows easier response and adaptation to environmental conditions (e.g. changes in network traffic or infrastructure). As such, they build the core of modern networks and are crucial to achieving the low latency and high speed of 5G networks. However, this makes VNFs of particular interest to cyber criminals, hacktivists, and state-sponsored hackers.

In October 2019, the EU Commission identified state-sponsored attackers as the major threat to the security of 5G networks. The EU's risk assessment identified core security requirements that are different for 5G networks. Due to the reliance on software, types of devices, and services connected and the heavily interconnected nature of 5G networks, there are more entry points for attackers. Nokia's head of product management security has said that 5G networks have 200 times more attack vectors than their 4G predecessors. Network services such as VIMs (virtualized infrastructure managers) have already been identified as crucial assets that are expected to be heavily attacked.

This article investigates how selected NATO and Western allies have addressed these issues of 5G network security over the past two years, while the pandemic has further highlighted societal dependency on network infrastructure. In particular, it will consider software supply chain security and the approach to foreign vendor integration. It investigates to what extent allies share views and practices on 5G security, which is necessary to ensure a united, secure network across borders. It will also consider the

implications of adversary activities directed against identified weaknesses and offer essential principles for how to cope with the emerging threats.

Keywords: *5G networks, software supply chain security, emerging threats, cyber conflict, cooperative and coordinated defence*

1. INTRODUCTION

Societies will soon be highly dependent on new and highly complex communications technologies currently being rolled out. But these technologies introduce a whole range of new vulnerabilities ripe for exploitation by both state and non-state actors. This paper considers the precautions that must be taken at a state and collaborative level, such as through NATO, in order to avoid worst-case scenarios where adversaries leverage these new vulnerabilities to achieve everything from low-level damage to strategic geopolitical gains.

Many of the application scenarios envisioned in the near future (such as IoT, autonomous driving, and Industry 4.0) will utilize technologies that are increasingly dependent on virtualizations to provide adequate and adaptable network services. These new technologies will determine the digital evolution of the coming years and have in common an increasingly complex underlying technical and strategic configuration. This will make it increasingly hard for decision-makers to estimate the impact their decisions have on the underlying technology and whether political decisions made are technically feasible.

A purely technical approach to the security and protection of critical network infrastructure will be insufficient to keep moving forward securely within the upcoming networks, given recent geopolitical developments. We will discuss how the technology shifts experienced will also impact strategic and political decision-making and how to cope with the implications this has.

5G networks are often perceived as a new networking standard that simply uses a number of new protocols and frequencies. But this is an oversimplified picture of the nature of 5G networks. The underlying structure consists of a mixture of soft- and hardware components whose complexity and orchestration may be difficult even for dedicated experts to grasp fully. This becomes even more evident when security predictions are to be made or when policy regulations are to be implemented technically. Given this level of complexity, it is difficult to lead an adequate, appropriate, and reality-

based discussion between technical and non-technical stakeholders. Attempting to do so typically leads to a high level of abstraction and a meta-language that hides any underlying technology. This abstract and veiled discussion does not provide decision-makers with enough knowledge about the underlying properties to make well-based decisions and – conversely – makes it harder for IT professionals to translate regulatory language into technical implementations, as the demands are technically too unspecific and abstract and sometimes even incompatible with the underlying architecture. This communicative difficulty is already a security threat in its own right and significantly exacerbates already complex technical security concerns.

In October 2019, the EU Commission identified state-sponsored attackers as the main threat to the security of 5G.[1] The EU's risk assessment identified core security requirements that are different for 5G networks. Due to the reliance on software, the types of devices and services connected, and the heavily interconnected nature of 5G networks, there are more entry points for attackers. Nokia's head of product management security stated that 5G networks have 200 times more attack vectors than their 4G predecessors.[2] Network services such as VIMs (virtualized infrastructure managers) have already been identified as crucial assets[3] that are expected to be attacked heavily.

The European Union Agency for Cybersecurity (ENISA) report 'Threat Landscape for 5G Networks'[4] describes a list of evolving threats, and multiple experts have published their findings and concerns regarding the security of the underlying technology. While one may assume that the technical community is aware of these threats (although the same assumption may not apply to anyone outside this largely self-contained bubble), what is truly lacking is a discussion on how these threats may impact political decisions and how adversary strategies may utilize this evolving threat landscape to their benefit. The prominent discussion that immediately comes to mind is the exclusion of Huawei from contributing to the development of 5G networks in a range of Western nations. Although this might suggest that there is already adequate discussion on 5G security concerns and their policy implications, closer investigation proves the contrary.

The concerns addressed in the Huawei exclusion were mainly regarding the development of hardware components. From the view of an IT security professional,

1 NIS Cooperation Group, 'EU Coordinated Risk Assessment of the Cybersecurity of 5G Networks', Report, 9 October 2019: 'This will, in turn, increase the number of attack paths that could be exploited by threat actors, in particular non-EU state or state-backed actors, because of their capabilities (intent and resources) to perform attacks against EU Member States telecommunications networks, as well as the potential severity of the impact of such attacks.'

2 Emily Jackson, '5G Has 200 Times More Access Points for Hackers than Existing Networks, Experts Warn', *Financial Post*, 24 January 2019, https://financialpost.com/telecom/attack-surface-has-multiplied-5g-networks-more-vulnerable-to-hackers-conference-told.

3 ENISA, 'Security Aspects of Virtualization', Report, February 2017.

4 ENISA, 'ENISA Threat Landscape for 5G Networks: Threat Assessment for the Fifth Generation of Mobile Telecommunications Networks (5G)', Report, November 2019.

hardware security flaws are, of course, not to be neglected, but the more serious concern in 5G networks ought to be around the software attack surface. This was not a prominent topic within the public discussion, nor did the discussion lead to clear guidance for how Western nations ought to proceed with Huawei or other software contributors that are considered to be controlled by adversarial state interests.

As 5G, and the already planned 6G, is currently forming the backbone of the networks we use for a multitude of applications, we need to ensure that security aspects are considered adequately and jointly on both the strategic and technical level. Technical flaws in future network backbones resulting from inadequate translation or communication between strategic decision-makers and technical practitioners may render Western nations susceptible to politico-technical blackmail by adversary actors. This is especially the case if the adversary has the technical knowledge and power to abuse these weaknesses and if the adversary is aware of these weaknesses before Western nations are. This is a one-way street – this weakness cannot be used by Western nations for retaliatory offensive measures against such adversaries.

Some of the crucial questions that must be asked among allies are: do we consider security in 5G networks equally, and do we take appropriate actions? Are these actions technically feasible, and do the technical implementations correspond to our strategic plans? Can we agree on equivalent levels of security to be met, and what are our measures to uphold this level of security? Are these measures compatible with each other, especially on the technical level? Or do they induce further security threats? And most fundamentally of all, have plans for networks provided for their protection and resilience against attacks using state-level capabilities and resources?

The remainder of this article investigates how NATO and Western allies have addressed these issues of 5G network security over the past years, while the pandemic has highlighted societal dependency on network infrastructure still further. In particular, it will consider software supply chain security and the approach to third-party contributions. It investigates to what extent allies share views and practices on 5G security, which is necessary to ensure a united, secure network across borders. It will also examine adversary activities directed against identified weaknesses and offer guidelines on how to cope with the emerging threats.

2. 5G SECURITY CONSIDERATIONS

A proper grasp of the new security challenges presented by 5G requires an understanding of the highly complex technology involved. 5G networking introduces a number of technologies that political and strategic decision-makers may be unaware of

but that are critical to understanding the risks incurred. These include the virtualization of hardware components, network slicing, software-defined networking, and the orchestration of these software components. All of these introduce CI/CD (continuous integration/continuous deployment) security aspects to network backbones, because the essential software will need to be developed and deployed continuously, opening up the backbones to CI/CD attack patterns.

CI/CD security is already an issue today, but usually this only affects software products – that is, a specific program, not the general backbone(s) of critical national infrastructure. Hence, CI/CD attacks have until now often only been relevant to software firms producing high-value targets, such as operating system developers. But in the same way as attacks against virtualized environments (considered further below), both the likelihood of attack and the profile of attackers are likely to change substantially due to the unprecedented nature of the 5G environment. An additional danger is that one attack pattern may be used to infiltrate the entire backbone; previously it was only able to attack a single software product. In effect, a single CI/CD attack against a suitable software component not only can be deployed against a wide geographical area but also, through a partially autonomous roll-out of updates, can easily and rapidly propagate to a wide variety of networks.

As ENISA states in its recent report on the 5G security landscape:[5]

> One of the most important innovations in the 5G architecture is the complete virtualisation of the Core network.... These novel network technologies and concepts that rely heavily on 'softwarisation' and virtualisation of network functions will introduce new and complex threats. The core network is the central part of the 5G infrastructure and enables all functions related to multi-access technologies. Its main purpose is to deliver services over all kinds of networks (wireless, fixed, converged).

A. Virtualization and Software-Based Architecture

At the core of the development of 5G lies the transition of several architectural components from previous hardware-based to software solutions. This transition has allowed the introduction of new functionalities that are crucial to meeting the demands that 5G networks were developed to address.

Virtualized network functions (VNFs) are used to provide services through software that replaces dedicated network devices. This shift from physical devices to software functions allows easier response and adaptation to environmental conditions (e.g. changes in network traffic or infrastructure). As such, they build the core of modern networks and are crucial to achieving the low latency and high speed of

5 ENISA, 'Threat Landscape for 5G Networks', Report, December 2020.

5G networks which are currently being deployed. They build the new backbone of our upcoming networks. However, this makes VNFs of particular interest to cyber criminals, hacktivists, and state-sponsored hackers, especially as new attack vectors are introduced.

While virtualization does provide some means of inherent security, especially against attacks found on the lower end of technical capabilities, it also introduces new vulnerabilities, especially when expecting malicious actors from the higher end – dedicated, state-sponsored IT professionals paid to invade and infiltrate target networks. The particular threats that arise in 5G networks based on virtualization involve the possibility of 'cross-contamination' of shared hardware resources, the introduction of software components as basic network functions lacking integrity, and in general, the attack vectors associated with virtualization environments (e.g. such as hypervisor vulnerabilities and virtual machine escape attacks). Currently, it is anticipated that 5G network security to cope with virtualization attacks needs to be standardized and may be integrated fully based on existing solutions. However, the vast examples of attack vectors directed against virtualizations do make this approach questionable. Many of these attack vectors have been considered rather theoretical attacks in the past, as they required skilled personnel and financial resources. It was presumed that non-state actors were rather unlikely to make such an investment, especially as valuable targets usually combined virtual and physical security measures. This argument is no longer valid, since we must expect primarily state- and state-sponsored actors to attack the backbone of 5G networks and can no longer rely on physical security in 5G.

The fact that attack vectors against virtual environments are – compared to regular attacks experienced on networks today – relatively complex to carry out renders them even more likely to appear on 5G networks, as the attacker profile is likely to differ from today's network attackers. This may change over time, but it is reasonable to expect that 5G networks will initially be a sector for high-skilled attackers rather than a playground for off-the-shelf exploits and 'script kiddies'.

NFVs (network function virtualizations, a structure used for technical management and orchestration) are used to manage and orchestrate the virtualized basic, previously hardware-based, network functions. In order to do so, software components need to be addressable (in technical terms) in a standardized way. This is realized through inter-component interfaces, known as APIs (application programming interfaces). These APIs are challenging to design without introducing security vulnerabilities for a vast variety of technical reasons, whose explanation would exceed this article's scope but which may be categorized as 'API vulnerabilities'. Other expected vulnerabilities are account compromise, unauthorized (privileged) user access, unauthorized inspection or modification of data, and compromise of management and orchestration (MANO)

components. This may lead to an invasion and subsequent control over the MANO module of the network and thus unauthorized control over all communication, functions, and operations performed through this network.[6]

B. Network Slicing

One of the achievements of the 'softwarization' of networks is the introduction of what are known as 'network slices'. Network slices allow the provision of ad-hoc networks perfectly fitting the current demands of the users. They are built on the integration of various software components, on demand, providing a specific network view. Basically, this concept is similar to the virtual end-to-end networks known today. However, in this case it is not merely a common channel being used by two parties but rather a set of network components (which are provided by means of software, such as VNF) that are orchestrated into a bundle, which may then be shared among a given set of users, meeting their specific demands at that time. This inherently introduces the need for adequate business and trust models in order to know whom to trust with the provision and exchange of components, as well as complex orchestration mechanisms to uphold the dedicated security level (see Section D). This introduction to slicing is necessarily brief, but more detailed explanations can be found in the references below.[7]

While authentication and authorization mechanisms have been introduced with release 16, several security concerns remain unaddressed. These concerns are particularly the handling of compromised or malfunctioning network slices and the security of shared media (recall that network slices are based on software components which run on third-party hardware – such conditions have been used in the past to compromise and extract information). Major security flaws have already been detected in the concept.[8]

C. Software Supply Chain

It is inherent to the architecture of 5G that its core is composed of software components jointly operated on shared hardware. These software components communicate among each other for coordination purposes (see above section on API security). Additionally, this yields the concern of software integrity and software supply chain security.[9] Software supply chain security is relevant not only for the initial

6 ENISA, 'Threat Landscape for 5G Networks', Section 3.7.2, 'Security Considerations', Report, December 2020.

7 Xenofon Foukas, Georgios Patounas, Ahmed Elmokashfi, and Mahesh K. Marina, 'Network Slicing in 5G: Survey and Challenges', IEEE Communications Magazine 55, no. 5 (May 2017), 94–100; NGMN Alliance. Description of network slicing concept. NGMN 5G P1, September 2016, https://www.ngmn.org/publications/description-of-network-slicing-concept.html.

8 Máirín O'Sullivan, '5G Network Slicing Vulnerability: Potential for Fraud or Data Leakage', AdaptiveMobile Security Blog, https://blog.adaptivemobile.com/5g-network-slicing-vulnerability-fraud-data-leakage; Vitor A. Cunha, Eduardo da Silva, Marcio B. de Carvalho, Daniel Corujo, Joao P. Barraca, Diogo Gomes, Lisandro Z. Granville, and Rui L. Aguiar, 'Network Slicing Security: Challenges and Directions', *Internet Technology Letters* 2, no. 5, September/October 2019 (first published 29 July 2019), https://doi.org/10.1002/itl2.125.

9 ENISA, 'Threat Landscape for 5G Networks', Section 3.3.2, 'Core Architecture (Zoom-in) – Security Considerations', December 2020.

development and deployment but for the whole software life cycle, especially for updating and maintenance purposes.

One of the goals of the 5G architecture was to allow various parties to easily contribute to the 5G networking landscape. The concept is that various network operators share the frequencies provided by the telecommunication provider, based on software and virtualized components. This inherently raises the question of the software's origin, integrity, and quality. Additionally, the vast attack surface of CI/CD security is introduced in order to allow continuous development and deployment.

Once installed, software needs to be maintained in order to respond to changes in the infrastructure, demands, or novel threat vectors and weaknesses. While hardware components have been affected by an increasingly short life cycle over the past decades, this life cycle still spans from months to years before a hardware component is considered outdated. Code, which produces the software product, has a much shorter life cycle, often spanning days to weeks. If hardware is considered static, modern software should be considered highly versatile and agile.

The 'softwarization' of the 5G core also leads to the necessity to keep these components updated and securely compatible, as a gap between different components may in turn again introduce new threats, lead to malfunctions, and reduce the overall security of the network. A rushed update (e.g. in order to swiftly respond to a novel attack vector observed during an attack) may weaken the overall network structure even further. The network's protection will only be as strong as its weakest software component, or the weakest link between software components. This software component may be shared and the sharing will often remain unnoticed from within the network, as well as the use of infiltrated software or background updates. This puts further responsibility on the MANO.

D. Management and Orchestration (MANO)[10]

The multitude of actors, domain environments, roles, and rights involved make the management and orchestration of network slicing a challenging task. This is even more true when considering the security implications. Technical flexibility comes at the price of higher complexity and security risks. It is to be expected that attackers will be aware of this fact and that attacks will be developed to directly target this weakness. While security and authorization requirements for management services have been defined, these have not yet been implemented.

Another important aspect is that of post-incident analysis within network slices. Measures must be taken to provide logs of the activities from within the network slice;

10 ENISA, 'Threat Landscape for 5G Networks', Section 3.5, 'Management and Network Orchestration (MANO)' and Section 4.5 'Vulnerability Groups for Network Function Virtualization – Mano', December 2020.

however, this automatically introduces the need for several unrelated slices to write to a common, shared medium, which is generally considered a security vulnerability in virtualizations. Yet if this information is not provided, network slicing may be used to convey malicious actors' activities, making attribution even more complex.

E. Actions Among Allies

Western nations vary widely in their approach to the security of 5G networks and in the basic assumptions that drive their policy. This is clearly illustrated by the example of Huawei and other foreign technology providers being excluded from the development of 5G in some Western nations. While Sweden has excluded the two Chinese contributors, Huawei and ZTE, from participating in the development of 5G in Sweden (with Huawei filing an appeal against this ban),[11] Norway has decided not to strictly ban the use of 5G equipment. Despite demands from the Norwegian government in December 2019, Telenor's development of the 5G network in Bergen, Norway, is mostly based on Huawei sources and will remain so until 2024, when Telenor will remove untrusted sources due to security considerations.[12] Austria initially decided not to rule out the use of Huawei technology in 5G networks but stated (under former chancellor Sebastian Kurz) that it would coordinate its actions with the EU.[13] A more recent statement indicates that 5G may be built with Chinese contributions in Austria, serving Telekom Austria's 25 million customers across Austria, Bulgaria, Croatia, Belarus, Slovenia, Serbia, and North Macedonia and using Chinese vendors in Bulgaria and North Macedonia for 4G.[14] Belgium initially found no evidence of possible espionage through the utilization of Huawei's technology and hence decided not to ban Huawei from contributions to the Belgian 5G network development.[15] However, it was reported that Belgian operators Orange Belgium and Proximus dropped Huawei as a consequence of US pressure to exclude Chinese vendors.[16] It was later reported that Belgium had previously experienced a pro-Huawei

11 Reuters, 'Huawei Appeals Sweden's Ban on Company for Selling 5G Gear', 1 October 2021, Reuters, https://www.reuters.com/business/media-telecom/huawei-appeals-swedens-ban-company-selling-5g-gear-2021-10-01/.

12 Gregers Møller, 'Experts: Huawei Equipment Used for Expansion in Bergen Is a Cause for Concern', *ScandAsia*, 8 October 2021, https://scandasia.com/experts-huawei-equipment-used-for-expansion-in-bergen-is-a-cause-for-concern/. 'Despite clear demands from the Norwegian government and several reports of concern regarding safety concerns related to the use of Chinese technology, the Chinese telecom giant Huawei is still responsible for most of the equipment used for the development of the 5G network in Bergen, Norway and that is a cause for concern. Experts believe that if they want, the technology can be controlled from China, *Bergens Tidende* writes quoting *Bergensavisen* and *Nettavisen*'.

13 Reuters Staff, 'Austria to Collaborate with EU Partners on Huawei 5G Decision', Reuters, 20 January 2020, https://www.reuters.com/article/us-austria-5g-huawei-tech/austria-to-collaborate-with-eu-partners-on-huawei-5g-decision-idUSKBN1ZJ10R.

14 Supantha Mukherjee, 'Telekom Austria May Consider Huawei, ZTE for 5G Networks: COO', Reuters, 23 June 2021, https://www.reuters.com/business/media-telecom/telekom-austria-may-consider-huawei-zte-5g-networks-coo-2021-06-23/.

15 Amée Zoutberg, 'Belgium Will Not Join UK in Banning Huawei from Its Telecom Networks', *Brussels Times*, 14 July 2020, https://www.brusselstimes.com/news/belgium-all-news/121568/belgium-will-not-join-uk-in-banning-huawei-from-its-telecom-networks.

16 Supantha Mukherjee and Mathieu Rosemain, 'Huawei Ousted from Heart of EU as Nokia Wins Belgian 5G Contracts', Reuters, 9 October 2020, https://www.reuters.com/article/us-orange-nokia-security-5g-idUSKBN26U0YY.

malign influence campaign.[17] Britain decided to remove Huawei's technology from its telecommunication networks and demanded its vendors reduce Huawei's share of the network infrastructure to 35% by 2023.[18] This process, which is expected to take until 2027, was instigated by the United States, alleging that Huawei posed a security threat due to its closeness to the Chinese government.[19] Orange France decided to avoid using Chinese vendors when developing European 5G networks but envisions Huawei playing a role in the African 5G rollout.[20] If there is a pattern among Western approaches to 5G security, it appears to be a consensus within the 'Five Eyes' partnership, and wide variance outside it.[21]

Many of the discussions around whether Huawei equipment (or software, as described above) should be used in 5G networks centred on the question of whether a 'backdoor' existed in available equipment. The crucial realization, however, is that the 5G architecture does not need 'backdoors' to be built into the system. The mere ability to contribute to the 5G core network constitutes a backdoor. The 5G architecture demands continuous integration and continuous deployment of function and software updates. As there is currently no valid option to automatically check code iterations for complex malicious execution options, this either must be monitored by IT professionals or will simply be based on trust in the contributor. The huge number of code iterations and updates that will propagate through a multitude of virtualizations makes it unlikely that this monitoring can be done by humans.

In addition, the complexity of the architecture and the distributed nature of its development opens up a still greater risk of supply chain attacks. In mid-2020, detection of the SolarWinds attack demonstrated the critical importance of software supply chain transparency and integrity (as well as the power of CI/CD attacks, as described above).[22] It is equally vital that the construction of 5G network architecture for NATO member states is protected from hostile state interference – but this begs

17 Adam Satariano, 'Inside a Pro-Huawei Influence Campaign: A Covert Online Push to Sway Telecommunications Policy in Favor of the Chinese Company May Presage a New Twist in Social Manipulation', *New York Times*, 29 January 2021, https://www.nytimes.com/2021/01/29/technology/commercial-disinformation-huawei-belgium.html.

18 Ryan Browne, 'British Mobile Carriers Warn Removing Huawei Will Cause "Blackouts" and Cost Billions', CNBC, 9 July 2020, https://www.cnbc.com/2020/07/09/vodafone-and-bt-warn-about-cost-disruption-of-removing-huawei.html.

19 Thomas Seal, 'BT's $700 Million Job to Rip-And-Replace Huawei 5G Begins Here', Bloomberg, 14 May 2021, https://www.bloomberg.com/news/articles/2021-05-14/bt-s-700-million-job-to-rip-and-replace-huawei-5g-begins-here.

20 Clara-Laeila Laudette and Supantha Mukherjee, 'Orange Sees Role for Huawei in 5G Africa Rollout', Reuters, 30 June 2021, https://www.reuters.com/business/media-telecom/orange-sees-role-huawei-5g-africa-rollout-2021-06-29/.

21 A broader overview of other nations' stances on the regulations of Chinese vendors' contributions to their 5G backbone deployment is available at Reuters Staff, 'Factbox: Huawei's Involvement in 5G Telecoms Networks around the World', Reuters, 20 October 2020, https://www.reuters.com/article/idUSL8N2GR34Y.

22 Lily Hay Newman, 'A Year After the SolarWinds Hack, Supply Chain Threats Still Loom', *Wired*, 12 August 2021, https://www.wired.com/story/solarwinds-hack-supply-chain-threats-improvements/.

the question of how to exclude software providers owned by interests within those adversary states if they are already integrated into the network provision ecosystem.[23]

3. STRATEGIC GOALS OF ADVERSARIES IN NETWORKS OF NATO ALLIES

It is essential that any new technology that introduces critical dependencies within NATO nations be adequately secured against threats from both non-state and state actors. This requires a full and regularly updated assessment of the aims and approaches to conflict of a wide range of possible adversaries. It also requires full and honest recognition of the threat among NATO allies themselves, and an acceptance that a state of notional peacetime does not mean that hostilities are not being waged by any available means.

The new vulnerabilities introduced by the specific nature of 5G networks lend themselves to a wide range of unfriendly and overtly hostile actions by adversaries. As with other forms of information threat, these span a broad spectrum of ambition, from simply causing damage with no other specific objective in mind, to high-level geostrategic change brought about through indirect means.[24] Non-like-minded nations, including but not limited to Russia and China, have closely studied means of damaging or destroying the civilian communications networks of NATO member states, and it should be anticipated that this probing for vulnerabilities will intensify as relations deteriorate further. Russian president Vladimir Putin has repeatedly promised responses to Western actions that are unexpected and 'asymmetric'.[25] To prevent such responses, NATO governments should seek to minimize their self-inflicted vulnerabilities.

Nevertheless, unless adequately secured, 5G networks simplify the task of the attacker in achieving their aims. Examples include:

Network destruction and information interdiction. Adversaries can achieve effects remotely that currently require physical intervention against telecommunications infrastructure.[26] This could be either in support of a localized objective, or a widespread attack in order to intimidate or blackmail victim states into political concessions.

23 See, for example, Brain4Net, a provider of software-defined networks (SDNs), NFVs, and more, listed on Intel's Network Builders site as part of their ecosystem but acquired by Kaspersky Labs in late 2021. 'Kaspersky Acquires Brain4Net', Channel Post Middle East and Africa, 7 November 2021, https://channelpostmea.com/2021/11/07/kaspersky-acquires-brain4net/.

24 *NATO Handbook of Russian Information Warfare*, NATO Defence College, November 2016, https://www.ndc.nato.int/news/news.php?icode=995.

25 Max Seddon and Henry Foy, 'Putin Threatens "Asymmetric" and "Tough" Response to US Sanctions', *Financial Times*, 21 April 2021, https://www.ft.com/content/724560a3-5b8e-483d-8262-e62053247366.

26 K. Giles and K. Hartmann, 'Adversary Targeting of Civilian Telecommunications Infrastructure', 2021 13th International Conference on Cyber Conflict (CyCon), 2021, 133–150, doi: 10.23919/CyCon51939.2021.9468303.

Infiltration, espionage, and situational awareness. The substantial increase in the number of attack surfaces that must be protected will facilitate attempts at stealthy penetration of networks for the purpose of long-term surveillance and data collection.

Subversion and other sub-threshold attacks. The nature of 5G networks will introduce an additional layer of deniability to attacks on communications networks and connectivity, as attribution becomes more technically, and thus especially politically, complex.

In addition, the nature of 5G, as well as other advanced technologies' reliance on it, presents adversaries with opportunities for the innovative exploitation of vulnerabilities. Cyber blackmail for political coercion is not a novel concept, but it takes on new dimensions thanks to the propagation characteristics of 5G. For demonstrations of the blackmailer's capabilities, destruction of critical national infrastructure remains a relatively unlikely option, given the near-consensus that this constitutes an act of war. However, if such an attack were carried out, isolation measures intended to protect critical infrastructure would limit its reach. By contrast, more subtle interventions – like increasing the latency of the virtual network functions used for autonomous driving – allow less detectable operations but, at the same time, could affect much wider areas and spread throughout broad networks. In other words, rather than attacking an element of critical infrastructure, such as a water supplier, attacking a NFV will hit not just that one water supplier but any network operator using the NFV. And depending on the NFV (or any other software component crucial to the backbone), this may propagate rapidly and widely, including across borders – like a domino at a central junction.

4. RECOMMENDATIONS

As described in Section 2, 5G gives rise to new technical attack vectors, as well as old ones. However, this fact is true for most new technologies, and it is not the reason for this article. The urgency arises due to the processes involved and the effects that the gap between technological implementation and strategic decision may have on the political level. It is important for decision-makers to realize that a basic understanding of the underlying technology is essential and cannot be delegated.

A. Coping with Novel Threats

The technologies used to build our future networks will not only allow us to move forward in application scenario terms but also force us to do so technically and strategically. As an example, ENISA reports that, while it was previously impossible to migrate the workload of a technical component from one service provider outside

of the defined legal and policy boundaries of that provider without notice, this threat is now technically feasible. In fact, as georestrictions cannot be enforced on NFVs in 5G, it is possible to move VNF from one location to another undetected.[27]

Consideration of some of the possible goals of adversary activities and the possible motivation for attacking 5G networks highlights that the threat should not be assessed only in terms of espionage and destruction. Infiltration, control, and strategic and political tactics should also be taken as serious concerns. 5G not only provides us with the capability to accelerate technologically but also provides our adversaries with a new dimension of operational domain. While cyberspace has always been the domain considered most easily attacked remotely and without attribution, 5G takes this statement to a new level. Through the softwarization of our networks, we remove a large portion of the last barrier of what has often been considered the 'physical security' of our network backbones. This is a development that cannot be stopped – in fact, reliance on this technology is firmly built into future development plans. But the nature of the threat, especially on the strategic and operational layer, dictates taking precautions now.

B. Exchange Among Professions

As the relevant European institution, ENISA does an excellent job of keeping experts updated and delivering reports on recent developments. However, these reports have a technical depth that will most probably make them incomprehensible to non-specialists. We strongly encourage lively exchange between technical and strategic personnel. An example of how this can be achieved is to bring together national and international decision-makers and developers in a similar manner as done in agile software development.[28] The objective of doing so would be bi-directional: to foster a deeper understanding of the underlying technology at the strategic level and, at the same time, to accelerate the understanding of strategic and political considerations and implications among technical personnel. This requires both groups to have the opportunity to easily and directly exchange their thoughts.

C. Consensus Strategy and Implementation

Furthermore, Western allies must ensure that they not only enforce strategic decisions nationally but also achieve common consensus policies. This will allow limitation in the variation of the technical implementations (which is already considerable) and may reduce incompatibilities. It also provides the chance to better control the gap between policy and implementation. Furthermore, strategies on adequate and secure CI/CD from a strategic view, especially on propagating software updates, must be

27 'Localisation of functions: Attacks aiming to place and migrate workload outside the legal boundaries were not possible using traditional infrastructure. Using NFV, violation of regulatory policies and laws becomes possible by moving one VNF from a legal location to another illegal location, because there is no mechanism to enforce georestrictions.' (ENISA Threat Landscape for 5G Networks).

28 Kim Hartmann, 'Shift Left: Secure by Design and Agile Development' (Original title: 'Linksruck – Shift Left – Secure by Design und agile Entwicklung', special issue, *iX* 16 (2020): 66–71.

developed and enforced. Measures must be installed to identify and remove malicious software instances among allies. Ways of ensuring this must be discussed in the context of allied and multinational operations and meeting different nations' demands collaboratively. Further research and discussion in this field is needed.

D. 'On-Boarding'

As technical implementation is the practical enforcement of whatever strategic decision is made, awareness among technical personnel of the strategic and political implications of technical decisions must be raised. Additionally, it is reasonable to demand that national institutions place decision-makers in the developer teams of software components that have such a wide impact on national cyber sovereignty.

An example of how this could be done is in the context of agile development by creating a dedicated role within the team. Due to the novelty of the technology and the beginning of an era where cyber sovereignty and the security of states are being placed in the hands of software developers, this role must urgently be defined for exactly this purpose – since it has not been previously required, it does not currently exist. While the exact outline of such a role and the professional skills required still need to be defined, we stress that current practice is inadequate for the described purposes, as it broadens the gap between policy-level stakeholders and the implementation team while not truly ensuring that someone on the team fully understands all technical, political, and strategic decisions.

This approach is feasible if the stakeholder is only interested in the final product outcome but does not need to understand how this product is achieved and if the developers do not care about the political or strategic implications of their development decisions (outside their technical scope). However, when it comes to the development of the backbone of our future networks, the strategic and political knowledge of decision-makers must be integrated into the development process, and strategic and political decision-makers must be aware of how their decisions are being implemented and the implications of the technical decisions being made. If it turns out to be impossible to empower decision-makers to this extent in terms of technical expertise, then consideration should be given to how dedicated professionals with expertise regarding the intersection of technical and political decisions can be educated. How this can be achieved or if there are other ways to solve these difficulties may be a subject for further research.

5. CONCLUSION

Future network technologies will be critical enablers for economic development and prosperity in NATO member states and among Western allies. But the steady deterioration of relations with key adversaries – and the emergence of new ones – means that any new capability must be carefully assessed to ensure that, along with prosperity, it does not bring a means to severely harm the interests of the state introducing it.

Compared with previous means by which adversaries could attack one or more NATO member states or Western allies, 5G network technology combines a unique, and potentially uniquely damaging, set of attributes. It will be ubiquitous; other key technologies, including critical national infrastructure, will be highly dependent on it; the attack surface is accelerated; and most importantly of all, it will be easily accessible to hostile actors once they establish themselves as contributors to the software backbone. The result is that as well as a radical step forward in telecommunications capability, 5G also risks offering an open door to those who would wish to cause harm.

It is therefore vitally important that planning for further rollout of 5G takes full account of the fact that it is being deployed in a hostile world. An essential element of this is a security mindset that combines geopolitical and technical awareness. This awareness must, in turn, inform a critical assessment of who should and should not be trusted to contribute to the construction of the networks – and, just as vitally, whom they, in turn, enlist as subcontractors. Given the complexity of the software backbone under construction today, it is unlikely that software from contributors found to be questionable can be fully removed at a later date. It is even unlikely that covert operations to compromise network components could be discovered at all. The implication is that NATO states may find themselves operating networks that are partly controlled by foreign nations without them even noticing, until it is too late – for instance, if a hostile power makes political demands while threatening to shut down public communications or network components serving critical infrastructure.

The other vital element of ensuring a secure architecture for 5G is ensuring that technical experts and government-level strategic planners are communicating clearly, effectively, and fully. The challenge of the language gap between technical and policy personnel is hardly a new one – but it becomes vastly more important as network architectures become increasingly incomprehensible to non-experts. Ensuring that both groups are talking to each other in a way that ensures that what is being discussed is implemented, and what is implemented is actually understood with all its strategic and political implications, is another essential safeguard against this vital new technology being used as a weapon against the states adopting it.