Published by CCDCOE

2020

12th International Conference on Cyber Conflict

20/20 Vision: The Next Decade

T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)

CYCON

IEEE
Advancing Technology for Humanity

2020
12TH INTERNATIONAL CONFERENCE ON CYBER CONFLICT
20/20 VISION: THE NEXT DECADE

IEEE Catalog Number: CFP2026N-PRT
ISBN (print): 978-9949-9904-6-7
ISBN (pdf): 978-9949-9904-7-4

COPYRIGHT AND REPRINT PERMISSIONS

NATO 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
Head of publishing: Jaanika Rannu
Layout: JDF

NATO COOPERATIVE CYBER DEFENCE CENTRE OF EXCELLENCE

NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) is an international military organisation, similar to other 24 NATO´s Centres of Excellence (COE). Although accredited by NATO, all COEs are independent research, training and exercise centres initiated by a Framework Nation (in the case of CCDCOE – Estonia) and belong to their member nations. COEs are not part of NATO command structure.

This Tallinn-based cyber defence facility conducts research, trainings and exercises in four core areas: technology, strategy, operations and law. The heart of the Centre is a diverse staff of international experts from military, government, academia and industry, currently representing 25 member nations from NATO Allies and like-minded partners beyond the Alliance, with many more on the path to joining. In short, CCDCOE is a NATO-accredited cyber defence hub that supports its member nations and NATO with cyber defence expertise, enabling a unique 360-degree approach to some of the most relevant cyber defence issues.

The Centre's training courses are based on the latest research and cyber defence exercises. The continuously updated selection of training courses addresses the emerging demands in the cyber defence. To best meet the training requirements of our Allies, Partners and NATO as a whole, The Centre provides courses in different formats and locations, including e-learning or training by a mobile team, covering a broad range of topics in the technical, legal, strategic and operational cyber security domains. Appointed by NATO as the Department Head for Cyberspace Operations Discipline, CCDCOE is responsible for identifying cyberspace operations training needs and matching them with education and training solutions for all NATO bodies across the Alliance.

Every spring the Centre hosts the annual conference on cyber conflict, CyCon, which unites decision-makers and experts from government, academia and industry from all over the world. In May 2019, CyCon brought to Tallinn around 600 cyber experts from more than 40 nations, the conference theme was 'Silent Battle'.

CCDCOE organises the largest and most complex international live-fire cyber defence exercise in the world – Locked Shields. The annual exercise enables cyber security experts to enhance their skills in defending national IT systems and critical infrastructure under real-time attacks. More than 1400 cyber experts from 30 nations took part in Locked Shields 2019.

The Centre is also home to the Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations, the most comprehensive analysis on how existing international law applies to operations in cyber space. It further hosts the Cyber Law Toolkit, an interactive web-based resource for legal professionals who work with matters at the intersection of international law and cyber operations. A regularly updated database of national cyber security documents, the International Cyber Developments Review (INCYDER) and comprehensive overviews of national cyber security organisations, together with other Centre's publications and research papers are available on the Centre's website.

TABLE OF CONTENTS

INTRODUCTION

Every edition of the CyCon proceedings is special. The book that you have just opened is particularly so: for the first time in CyCon's history, it stands alone with no physical conference accompanying it. The event that has, over the twelve years of its existence, become a staple item on the agenda of many a cyber enthusiast in public administration, military and academia, drawing over 600 participants each May to the Estonian capital, had to cede to SARS-CoV-2 and public health concerns.

True to the cyber ethos of thinking in categories of abstract and invisible threats, and in order to celebrate the work of our authors, CCDCOE has decided to carry on with the publication of the CyCon proceedings nonetheless. Moreover, at a time when a substantial part of professional and personal activities are moving online and the whole world struggles to maintain normality under unprecedented circumstances, the CyCon theme, 20/20 Vision: The Next Decade, could not be more relevant. More than ever we need to be aware of the new technologies, policies and legal frameworks that will shape the future at societal and personal levels, and we must ask how to best protect our values while ensuring that cyberspace becomes more transparent, predictable and safe.

As usual, the papers gathered in this book reflect the three CyCon tracks: technical, strategic and legal. Of the total of 19 papers appearing in the book, five cover technical topics, seven touch upon strategic and seven upon legal issues. There is a variety of topics but a trend of shifting paradigms in their respective disciplines can be discerned. With the development of cyberspace as a domain of operations, new ideas appear, but old concepts are also being revamped and their application to cyberspace tested. Papers discussing artificial intelligence, autonomous weapons or cloud services stand shoulder to shoulder with papers dealing with energy distribution networks, security of industrial control systems (ICS) or responses available to states under international law.

Jeff Kosseff thus explores retorsion as a possible response to malign activities in cyberspace, as does **Przemysław Roguski** with the hot topic of collective countermeasures. The latter author's second paper, co-authored with **Neal Kushwaha** and **Bruce W. Watson**, compares various national approaches towards storage of governmental data in cloud and identifies possible risks in entrusting this job to commercial service providers incorporated in foreign jurisdictions. Also on the legal track, **Livinus Nweke** and **Stephen Wolthusen** examine the regulatory frameworks related to the protection of personal data and their impact on sharing of threat information among critical infrastructure operators. Two articles touch upon the obligations of states in regard to the development and use of cyber weapons.

Aleksi Kajander, **Agnes Kasper** and **Evhen Tsybulenko** contend that the legal review of new weapons under Article 36 of Additional Protocol I to the Geneva Conventions is limited in its reach, and examine instead the positive obligation to ensure respect for the Conventions under Common Article 1, with a particular focus on autonomous weapons systems. In their turn, **Ivana Kudláčková**, **Jakub Harašta** and **David Wallace**, while also acknowledging the limitations of Article 36, see policy benefits in extending the legal review to software used in operations under the threshold of use of force. Last but not least, **Tina J. Park** and **Michael Switzer** offer a new perspective of the responsibility-to-protect norm and explore its applicability in cyberspace.

On the strategic track, the focus has been on military cyber operations and cyber conflict in general. There are two geographically focused papers; the one by **Bilyana Lilly** and **Joe Cheravitch** offers a comprehensive overview of the evolution of Russia's posture in information warfare, while the other, authored by **Frédérick Douzet**, **Louis Pétiniaud**, **Loqman Salamatian**, **Kevin Limonier**, **Kavé Salamatian** and **Thibaut Alchus**, discusses fragmentation of the Internet with a case study of border gateway protocol manipulations during the political crisis in Ukraine. **Matthias Schulze** takes a closer look at the use of cyber capabilities in conflict situations, examining it at the operational, tactical and strategic levels. **Martin C. Libicki** explores the implications of spill-over of a conflict in cyberspace into physical domains. **Christopher Whyte** adds artificial intelligence to the concoction and studies how the new technologies augment offensive cyber operations and what this can mean for states' deterrence policies. In a similar vein, **Keir Giles** and **Kim Hartmann** examine the impact of machine-learning on the execution of malign influence campaigns. Closing the strategic track, **Jason Healey**, **JD Work** and **Neil Jenkins**, through selected case studies, analyse how defenders have sought to disrupt adversary operations in cyberspace, offer an analytical framework to categorise such campaigns and measure their impact, while providing a unique dataset spanning over the last thirty years.

The topics of the five technical papers span from industrial control systems through artificial intelligence to post quantum cryptography. **Michael Dodson**, **Alastair R. Beresford** and **Mikael Vingaard** present a study of high-interaction ICS honeypots and argue that networks of Internet-connected honeypots can effectively be used to identify targeted ICS attacks in order to better defend systems that are known for their heterogeneity rendering a uniform approach difficult. **Gilberto Pires de Azevedo**, **Maxli Barroso Campos** and **Paulo César Pellanda** take the example of electric power systems and examine, from the cybersecurity perspective, their traditional structure and the foreseeable changes due to a convergence of environmental factors and the advent of new technologies, discussing how to mitigate the associated risks. **Artūrs Lavrenovs**, **Roman Graf** and **Kimmo Heinäaro** propose, in their paper,

utilising neural networks for automated classification of individual devices connected to the Internet and examine how to use HTTP features to train such networks. **Kim Hartmann** and **Christoph Steup** report on attack patterns directed against artificial intelligence and machine learning methods, which are likely to occur more frequently given society's increasing dependence on new technologies, and contemplate related policy considerations. The book concludes with a specifically focused paper on isogeny-based post-quantum cryptography authored by **Lubjana Beshaj** and **Andrew O. Hall** of the US Army Cyber Institute at West Point, our partner institution and organiser of the CyCon US conference.

All articles published in the book have been subjected to a double-blind peer review by at least two members of the CyCon Academic Review Committee. We are indebted to the reviewers who have invested their time and expertise to help us make the final selection.

We equally remain grateful to our authors and researchers who have chosen CyCon over other platforms to present their original work and who decided not to withdraw their papers although they had been deprived of the opportunity to present them and put their conclusions to the test in front of CyCon's conference participants.

Within this context, we want to particularly thank the Institute of Electrical and Electronic Engineers (IEEE) and its Estonian section for their continued support and technical sponsorship of the CyCon publications and for showing flexibility in uncertain times.

It goes without saying that the job would have only been half-done without the patient and often invisible work of CCDCOE staff, whom we thank for their efforts in first preparing for the CyCon conference and this book, and subsequently adapting to the new circumstances. Our gratitude namely goes (in alphabetical order) to Annika Kvelstein, Liis Poolak and Jaanika Rannu of CCDCOE's Support Branch for logistics support and to Henrik Beckvard, Costel-Marius Gheorghevici, Kadri Kaska, Liina Lumiste, Piret Pernik and Ann Väljataga for invaluable editor assistance.

THE EDITORS

Academic Review Committee Members for CyCon 2020:

- Dr Ali Hilal Al-Bayatti, De Montfort University, United Kingdom
- Siim Alatalu, Information System Authority, Estonia
- Geert Alberghs, Ministry of Defence, Belgium
- Lt.Col. Vincent Banse, NATO CCDCOE
- Lt.Col. Henrik Paludan Beckvard, NATO CCDCOE
- Lt.Col. Fabio Biondi, NATO CCDCOE
- Dr Bernhards Blumbergs, CERT.LV, NATO CCDCOE Ambassador, Latvia
- Maj. Pascal Brangetto, Ministry of Armed Forces, France
- Dr Joe Burton, Waikato University, New Zealand
- Prof Thomas Chen, University of London, United Kingdom
- Prof Michele Colajanni, University of Modena and Reggio Emilia, Italy
- Prof Didier Danet, Military Academy of Saint-Cyr, France
- Dr Thibault Debatty, Royal Military Academy, Belgium
- Prof Dorothy E. Denning, Naval Postgraduate School, United States
- Samuele De Tomas Colatin, NATO CCDCOE
- Prof Sybe De Vries, Utrecht University, Netherlands
- Prof Frédérick Douzet, GEODE; University Paris 8, France
- Dr Helen Eenmaa-Dimitrieva, University of Tartu, Estonia
- Dr Kenneth Geers, COMODO, NATO CCDCOE Ambassador, United States
- Prof Solange Ghernaoutti, INFORGE Institute, Lausanne University, Switzerland
- Capt. Costel-Marius Gheorghevici, NATO CCDCOE
- Keir Giles, Chatham House, Conflict Studies Research Centre, United Kingdom
- Prof Michael Grimaila, Air Force Institute of Technology, United States
- Dr Jonas Hallberg, Swedish Defence Research Agency, Sweden
- Dr Jakub Harašta, Masaryk University, Czech Republic
- Jason Healey, University of Columbia, United States
- Dr Trey Herr, Atlantic Council, United States
- Prof David Hutchison, Lancaster University, United Kingdom
- Dr Emma Irving, Leiden University, Netherlands
- Prof Gabriel Jakobson, CyberGem Consulting, United States
- Raik Jakschis, Hanse Digital Access, Germany
- Taťána Jančárková, NATO CCDCOE
- Dr Károly Kassai, Military National Security Service, Hungary
- Kadri Kaska, NATO CCDCOE
- Prof Sokratis K. Katsikas, Norwegian University of Science and Technology
- Prof Mika Kerttunen, Tallinn University of Technology, Estonia
- Dr Panagiotis Kikiras, European Defence Agency

- Assoc Prof Dan Dongseong Kim, University of Queensland, Australia
- Joonsoo Kim, National Security Research Institute, South Korea
- Dr Keiko Kono, NATO CCDCOE
- Markus Kont, NATO CCDCOE
- Dr Csaba Krasznay, National University of Public Service, Hungary
- Capt. Juha Kukkola, Finnish National Defence University, Finland
- Lt.Col. Aurimas Kuprys, NATO CCDCOE
- Lt.Col. Franz Lantenhammer, NATO CCDCOE
- Artūrs Lavrenovs, NATO CCDCOE
- Prof Sean Lawson, University of Utah, United States
- Ivan Lee, Singapore University of Technology and Design, Singapore
- Anne-Sophie Leonard, NATO CCDCOE
- Dr Lauri Lindström, NATO CCDCOE
- Liina Lumiste, NATO CCDCOE
- Dr Kubo Mačák, International Committee of the Red Cross, Switzerland
- Prof Olaf Manuel Maennel, Tallinn University of Technology, Estonia
- Dr Matti Mantere, Luminor Group, Estonia
- Prof Evangelos Markatos, University of Crete, Institute of Computer Science, Greece
- Prof Aditya Mathur, Singapore University of Technology and Design, Singapore
- Dr Paul Maxwell, Army Cyber Institute, United States
- Dr Stefano Mele, Italian Atlantic Committee, Italy
- Tomáš Minárik, NÚKIB, Czech Republic
- Dr Anna Molnár, National University of Public Service, Hungary
- Elsa Neeme, NATO CCDCOE
- Dr Lars Nicander, Swedish Defence University, Sweden
- Dr Julien Nocetti, IFRI, France
- Lt.Col. Gry-Mona Nordli, NATO CCDCOE
- Maj. Erwin Orye, NATO CCDCOE
- Dr Anna-Maria Osula, Guardtime, Tallinn University of Technology, Estonia
- Nicolas Ott, Organisation for Security and Co-operation in Europe
- Capt. (N) Barış Egemen Özkan, Ministry of National Defence, Turkey
- Dr Piroska Páll-Orosz, Ministry of Defence, Hungary
- Piret Pernik, NATO CCDCOE
- Mauno Pihelgas, NATO CCDCOE
- Prof Emanuela Pistoia, University of Teramo, Italy
- Capt. Roy Ragsdale, Army Cyber Institute, United States
- Dr Narasimha Reddy, Texas A&M University, United States
- Henry Rõigas, Guardtime, Estonia
- Prof Juha Röning, University of Oulu, Finland

- Lt.Col. Kurt Sanger, Department of Defense, United States
- Ragnhild Siedler, Norwegian Defence Research Establishment, Norway
- Lt.Col. Dr Massimilano Signoretti, NATO CCDCOE
- Dr Max Smeets, ETH Zurich, Switzerland
- Prof Edward Sobiesk, Army Cyber Institute, United States
- Dr Daniel Spiekermann, FernUni Hagen; German Police Forces, Germany
- Dr Tim Stevens, King's College London, United Kingdom
- Morta Strazdaite, Lithuanian National Cyber Security Center, Lithuania
- Dr Thierry Tardy, NATO Defence College
- Lt. (N) Ihsan Tolga, NATO CCDCOE
- Maria Tolppa, NATO CCDCOE
- Dr Jens Tölle, Fraunhofer FKIE, Germany
- Prof Risto Vaarandi, Tallinn University of Technology, Estonia
- Ann Väljataga, NATO CCDCOE
- Dr Adrian Venables, Tallinn University of Technology, Estonia
- Prof Ari Visa, Tampere University of Technology, Finland
- Lt.Col. Mark A. Visger, Army Cyber Institute, United States
- Maj. Gábor Visky, NATO CCDCOE
- Prof Col. David Wallace, U. S. Military Academy at West Point, United States
- Prof Bruce W. Watson, IP Blox, Stellenbosch University, South Africa
- Prof Sean Watts, Creighton University Law School, United States
- Cdr. (N) Michael Widmann, NATO CCDCOE
- Jan Wünsche, NATO CCDCOE
- Prof Stefano Zanero, Polytechnic University of Milan, Italy

CyCon 2020 Programme Committee:

- Dr Lauri Lindström, chair
- Lt.Col. Dr Massimiliano Signoretti, co-chair
- Lt. (N) Ihsan Tolga, co-chair
- Maj. Gábor Visky, co-chair
- Lt.Col. Henrik Paludan Beckvard
- Lt.Col. Fabio Biondi
- Capt. Costel-Marius Gheorghevici
- Taťána Jančárková
- Liina Lumiste
- Lt.Col. Gry-Mona Nordli
- Piret Pernik

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Retorsion as a Response to Ongoing Malign Cyber Operations

Jeff Kosseff
Assistant Professor
Cyber Science Department
United States Naval Academy[1]
Annapolis, MD
kosseff@usna.edu

Abstract: If a state has experienced a malicious cyber act that violates international law, it may implement proportional and limited countermeasures. If the act constitutes an armed attack, the target state may engage in self-defence. But what if the initial act, while malicious and harmful, does not clearly violate an international legal obligation? In such an instance, the primary option for response is retorsion, which is defined as an unfriendly but legal act. Little scholarship has meaningfully examined the contours of retorsion, which is increasingly important in an era of persistent, low-intensity cyber aggression. This paper seeks to fill that gap by exploring the contours of retorsion and examining the types of responses that could fall within its scope. It argues for an expansive view of retorsion that encompasses any responses that comport with international law. Definitional clarity is increasingly important to allow states to understand the range of potential responses to persistent cyber aggression that do not necessarily violate international law. Among the types of activities that may fall within the scope of retorsion are: exerting pressure via international relations, gathering information from the adversary's networks, observing the adversary on one's own network using tools such as honeypots, sending warnings to individual operatives, establishing a position on the adversary's systems and slowing down malign cyber operations.

Keywords: *retorsion, countermeasures, sovereignty, cyber*

1 The views expressed in this paper are only those of the author, and do not represent those of the United States Naval Academy, United States Department of Navy, United States Department of Defense, or any other party. Thanks to those who provided incredibly valuable feedback on earlier drafts, including Dennis Dias, Chris Inglis, Martin Libicki, and Kurt Sanger.

1. INTRODUCTION

In July 2019, the Netherlands Minister of Foreign Affairs released a nine-page summary of the government's views on international law as it applies to cyberspace. The document concluded with a discussion of states' response options, and much of the discussion focused on responses that have been thoroughly discussed in international law circles: countermeasures, pleas of necessity and self-defence. The document also highlighted a response that has not often been discussed in depth: retorsion.

As the Netherlands government described it, retorsion "relates to acts that, while unfriendly, are not in violation of international law".[2] Because retorsion is legal, the government wrote, it "is therefore always available to states that wish to respond to undesirable conduct by another state, because it is a lawful exercise of a state's sovereign powers."[3] The government listed a few examples: economic penalties, expelling diplomats, and "limiting or cutting off the other state's access to servers or other digital infrastructure in its territory".[4] Although the document only devoted two paragraphs to retorsion, the mere fact that a government highlighted it as one of the primary responses to malign cyber actions was noteworthy.

Retorsion is both flexible and limited. It is flexible because, unlike other responses, it is subject to relatively few operational requirements. It is limited because it may only consist of actions that comply with international law.

This paper highlights the reasons to classify a response to malign cyber activity as retorsion rather than countermeasures, and examines the types of responses that could qualify as retorsion. It argues for policymakers to broadly conceive of retorsion by including any responses – no matter how unfriendly – that comport with international legal norms, regardless of the legal status of the adversary's actions. A number of responses do not violate sovereignty or other international law. Very little scholarship has focused substantially on the boundaries of retorsion; this paper seeks to fill that gap.

A clearer understanding of retorsion is particularly useful as states confront persistent levels of cyber aggression that are below the level of an armed attack,[5] removing

2 Netherlands Minister of Foreign Affairs, Letter to the Parliament on the International Legal Order in Cyberspace (July 5, 2019), Appendix: International Law in Cyberspace, available at https://www.government.nl/binaries/government/documents/parliamentary-documents/2019/09/26/letter-to-the-parliament-on-the-international-legal-order-in-cyberspace/International+Law+in+the+Cyberdomain+-+Netherlands.pdf, at 7.

3 Ibid.

4 Ibid.

5 Statement of General Paul M. Nakasone, Commander, United States Cyber Command, Before the Senate Committee on Armed Services (Feb. 14, 2019), available at https://www.armed-services.senate.gov/imo/media/doc/Nakasone_02-14-19.pdf, at 2 ("The nation faces threats from a variety of malicious cyber actors, including non-state and criminal organisations, states, and their proxies. We see near-peer competitors conducting sustained campaigns below the level of armed conflict to erode American strength and gain strategic advantage").

self-defence as a potential response. Countermeasures will likely be available as a response to some of this aggression, but the use of countermeasures faces a number of constraints, described below. Many of the more aggressive responses will constitute countermeasures or even self-defence, but retorsion provides states with a flexible framework to respond to this persistent, low-level aggression.

2. THE LIMITS OF COUNTERMEASURES AS A LEGAL BASIS FOR RESPONSE

Countermeasures and retorsion are the primary legal categories of responses to cyber aggression that falls below the level of an armed attack. The main difference between them is that countermeasures would violate international law absent illegal actions by the adverse party,[6] while retorsion comports with international law regardless of the adverse party's actions.[7] Thus, outlining the limits that international law places on countermeasures helps to illustrate why it is useful to have a better understanding of the scope of retorsion.

Perhaps the most significant limit of countermeasures is that they only can be taken against a state that has violated an international legal obligation owed to the state seeking to take the countermeasure.[8] When it is debatable whether such a violation has occurred, there is uncertainty as to whether countermeasures are permissible. As one example, hacking that interferes with a state's electoral process could be viewed as a violation of the principle of non-intervention, which "forbids all States or groups of States to intervene directly or indirectly in internal or external affairs of other States",[9] but some commentators argue that election interference does not meet the standard for an illegal intervention.[10]

Often, such alleged violations that give rise to countermeasures involve breaches of sovereignty.[11] Determining whether such a breach has occurred in cyberspace is

6 Int'l Law Comm'n, Draft Articles on Responsibility of States for Internationally Wrongful Acts, Rep. of the Int'l Law Comm'n on the Work of Its Fifty-Third Session, U.N. Doc. A/56/10, at 75 (2001) [hereinafter ILC Draft Articles on Responsibility] at 75 ("In certain circumstances, the commission by one state of an internationally wrongful act may justify another state injured by that act in taking non-forcible countermeasures in order to procure its cessation and to achieve reparation for the injury").

7 Ibid. at 128.

8 Ibid. at 129 (Article 49.1) ("An injured State may only take countermeasures against a State which is responsible for an internationally wrongful act in order to induce that State to comply with its obligations under Part Two").

9 Craig Forcese, *The 'Hacked' US Election: Is International Law Silent, Faced with the Clatter of Cyrillic Keyboards*, JustSecurity (Dec. 16, 2016) , available at https://www.justsecurity.org/35652/hacked-election-international-law-silent-faced-clatter-cyrillic-keyboards (quoting *United States v. Nicaragua*).

10 Jens David Ohlin, *Did Russian Cyber Interference in the 2016 Election Violate International Law?* 95 Tex. L. Rev. 1579, 1587 (2017) (asserting that "the technical requirements for an illegal intervention might not apply to the Russian intervention, depending on how one understands the concept of coercion").

11 Michael N. Schmitt, *'Below the Threshold' Cyber Operations: The Countermeasures Response Option and International Law*, 54 Va. J. Intl'l L. 697, 704 (2014) ("In the cyber context, sovereignty grants a State the right (and in some cases the obligation) to regulate and control cyber activities and infrastructure on its territory").

often difficult, as there is no clear consensus as to whether an act of cyber aggression could constitute a standalone violation of sovereignty, or if it must implicate another rule such as non-intervention. The authors of the *Tallinn Manual* adopted the former view, writing that "[s]tates enjoy sovereignty over any cyber infrastructure located on their territory and activities associated with that cyber infrastructure".[12] Rule 4 of the *Tallinn Manual* provides that "[a] State must not conduct cyber operations that violate the sovereignty of another State".[13] This is consistent with the view of cyberspace sovereignty that the French Ministry of Armies released in 2019,[14] as well as that of some scholars.[15] In contrast, Jeremy Wright, then the Attorney General of the United Kingdom, said in 2018 that he was "not persuaded that we can currently extrapolate from that general principle a specific rule or additional prohibition for cyber activity beyond that of a prohibited intervention".[16] In other words, the general principle of sovereignty, in his view, did not create a bright-line rule that would be violated merely by virtue of an intrusion on the cyber infrastructure of another state. Similarly, an internal 2017 memorandum from the United States Department of Defense General Counsel stated that "[m]ilitary cyber activities that are neither a use of force, nor that violate the principle of non-intervention are largely unregulated by international law at this time".[17] Accordingly, states seeking to enact countermeasures may lack certainty that other states would view their actions as permissible responses.

In addition to the legal uncertainty over whether a particular act legally justifies countermeasures, the target state must have sufficient *factual* certainty of the source of the malign activity before engaging in countermeasures. As Michael Schmitt has noted, if its assessment of the origins of an attack "turns out not to be well-founded, the injured state's action cannot qualify as a countermeasure".[18] The Draft Articles on State Responsibility suggest "reasonable certainty" in attribution, leading Schmitt to conclude that "[a] cyber countermeasure undertaken in a mistaken but reasonable belief as to the identity of the originator or place of origin will be lawful so long as

12 *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*, at 11 (2017) [hereinafter Tallinn Manual].

13 Ibid. at 17.

14 Przemyslaw Roguski, *France's Declaration on International Law in Cyberspace: The Law of Peacetime Cyber Operations, Part I*, Opinio Juris (Sept. 24, 2019) ("From this France concludes that any cyberattack, i.e. any operation which breaches the confidentiality, integrity or availability of the targeted system, constitutes at minimum a violation of French sovereignty, if attributable to another State.").

15 Sean Watts & Theodore Richard, *Baseline Territorial Sovereignty and Cyberspace*, 22 Lewis & Clark L. Rev. 803, 808 (2018) (arguing that "the baseline rules of territorial sovereignty should be currently understood as a rule of conduct that generally prohibits states' nonconsensual interference with the integrity of cyber infrastructure on the territory of other states").

16 Speech, Rt. Hon. Jeremy Wright, *Cyber and International Law in the 21st Century* (May 23, 2018) , available at https://www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century ("Online as well as everywhere else, the principle of sovereignty should not be used by states to undermine fundamental rights and freedoms and the right balance must be struck between national security and the protection of privacy and human rights").

17 Watts & Richard (supra n 15) at 860 (quoting Memorandum from Jennifer M. O'Connor, Gen. Counsel of the Dep't of Def., International Law Framework for Employing Cyber Capabilities in Military Operations (Jan. 19, 2017)).

18 Schmitt (supra n 11) at 726.

all other requirements for countermeasures have been met".[19] Even under a flexible standard of "reasonable certainty," it may be difficult to attribute a malign act to a particular state.[20]

If a target nation identifies a violation of an international obligation and attributes it to another state with sufficient certainty, that state may engage in countermeasures; however, these countermeasures must be limited in purpose. Article 49 of the Draft Articles on Responsibility states that countermeasures may only be taken "to induce that state to comply with its obligations".[21] The purpose of limiting countermeasures is to reduce the likelihood of conflict escalation.[22] Article 49 limits countermeasures "for the time being",[23] which the drafters stated is meant to indicate "the temporary or provisional character of countermeasures".[24] If the initial malign actions that triggered the countermeasures are no longer occurring, the target state may no longer have the authority to engage in the countermeasures. Determining when illegal behaviour has ceased is difficult in cyberspace, particularly in light of the barrage of threats that nations face, often from the same handful of bad actors.

In addition to limits on the purpose and duration of countermeasures, international law restricts their magnitude. Article 51 of the Draft Articles requires that countermeasures be proportional, which means that they "must be commensurate with the injury suffered, taking into account the gravity of the internationally wrongful act and the rights in question".[25] The drafters of the *Tallinn Manual* suggested that when states consider whether countermeasures are proportional, they should assess "the injury suffered (i.e., the extent of harm), the gravity of the wrongful act (i.e., the significance of the primary rule breached), the rights of the injured and responsible State (and interests of other States that are affected, and the need to effectively cause the responsible State to comply with its obligations".[26]

Imagine that State A's government computer systems were taken offline for a day by a DDOS attack originating in State B. Using a countermeasure, State A might seek to cause damage to the State B computers that executed the attack. It would not be proportional, however, for State A to disable the power grid in an entire metropolitan area within State B.

19 Ibid. at 727.
20 Matthew C. Waxman, *Cyber-Attacks and the Use of Force: Back to the Future of Article 2(4)*, 36 Yale J. Int'l L. 421, 443 (2011) ("As a technical matter, those who study the problem of legally regulating cyber-attacks are usually quick to point out the problems of identification and attribution: it is not always possible to discern quickly or accurately who launched or directed an attack").
21 ILC Draft Articles on Responsibility at 129.
22 *Tallinn Manual* at 116.
23 ILC Draft Articles on Responsibility at 129.
24 Ibid.
25 Ibid. at 134.
26 *Tallinn Manual* at 128.

Countermeasures also face procedural restrictions. Most notably, a state engaging in countermeasures must "notify the responsible State of any decision to take countermeasures and offer to negotiate with that State [… though] the injured State may take such urgent countermeasures as are necessary to preserve its rights".[27] The notification requirement would pose little problem if the countermeasure was intended to dissuade the responsible state from continuing its malign acts. However, if the countermeasure was a cyber operation targeting the responsible state's systems that were responsible for the acts, notification would likely undercut the efficacy of the operation by providing a warning.

International law is also unsettled as to whether non-injured states may collectively engage in countermeasures on behalf of other states that are injured. In 2019, President Kersti Kaljulaid of Estonia took the position that international law allows collective countermeasures,[28] but that position at the moment is not widely accepted.[29] Unless there is a stronger international consensus that collective countermeasures are permissible, states will likely lack the necessary assurances to collaborate on countermeasures on behalf of an injured state.

In sum, countermeasures can be a useful tool to respond to low-intensity aggression in cyberspace, but their implementation is subject to many constraints. As states look to respond to this persistent malicious behaviour, they should consider whether some responses can be classified as retorsion rather than countermeasures.

3. RETORSION'S FLEXIBILITY

Countermeasures are subject to numerous restrictions because, absent the illegal acts of the responsible state, they would violate international law. If the actions underlying the countermeasures would not violate international law regardless of the actions of the other state, then it is unnecessary to classify them as countermeasures. Legal actions are retorsion that are not subject to the same limits as countermeasures. This section outlines the scope of retorsion and argues that it allows for a flexible approach to any unfriendly actions that comply with international law.

27 ILC Draft Articles on Responsibility at 135.

28 *'President Kaljulaid at CyCon 2019: Cyber Attacks Should Not be an Easy Weapon'*, ERR NEWS (May, 29, 2019), https://news.err.ee/946827/president-kaljulaid-at-cycon-2019-cyber-attacks-should-not-be-easy-weapon ("Estonia is furthering the position that states which are not directly injured may apply countermeasures to support the state directly affected by the malicious cyber operation").

29 Michael Schmitt, *"France's Major Statement on International Law and Cyber: An Assessment"*, JustSecurity (Sept. 16, 2019), available at https://www.justsecurity.org/66194/frances-major-statement-on-international-law-and-cyber-an-assessment/ ("Somewhat surprisingly in light of its central place in the NATO alliance and its key role in European security affairs, France rejects the position recently set forth by Estonian President Kersti Kaljulaid that collective countermeasures – that is, countermeasures taken by one State on behalf of another State that is entitled to take countermeasures by virtue of being the target of an unlawful cyber operation – are permissible").

Only a limited body of scholarship and jurisprudence has attempted to define retorsion, and often in a fleeting manner. The Draft Articles commentary describes retorsion as "'unfriendly' conduct which is not inconsistent with any international obligation of the State engaging in it even though it may be a response to an internationally wrongful act".[30] Likewise, the US Defense Department *Law of War Manual* characterises retorsion as "unfriendly conduct, (1) which is not inconsistent with any international obligation of the State engaging in it, and (2) which is done in response to an internationally wrongful act".[31] Leading treatises[32] and scholarship[33] similarly describe retorsion as unfriendly but legal actions. The scholarship typically focuses on diplomatic and economic forms of retorsion, such as sanctions.[34] However, if one were to take a broader view of retorsion so that it encompasses any unfriendly but legal response, these are but one form of retorsion.

A state that classifies its response as retorsion faces fewer legal constraints than if it employs countermeasures. The primary *legal* limit on retorsion is that, regardless of the actions of the adversary, it may not violate international legal obligations owed to other states. This eliminates a number of more aggressive options that may violate legal obligations,[35] but once a state has addressed the threshold concern of legality, it does not face as many legal restrictions on the purpose, duration, and character as a state employing countermeasures. Retorsion "casts a political shadow over the relationship between the two states", but such political effects are not legally prohibited.[36] Pragmatic concerns may limit retorsion, but such limits are not imposed by international law.

Unlike countermeasures, retorsion is not limited to responding to internationally wrongful acts. It may also be exercised in response to the unfriendly but legal acts

30 ILC Draft Articles on Responsibility at 128.

31 US Department of Defence, *Law of War Manual* (June 2015, Updated December 2016) at 1110 [hereinafter *Law of War Manual*].

32 Anthony Cassese, *International Law* (2d ed. 2005) at 310 ("Retortion embraces any retaliatory act by which a state responds, by an unfriendly act not amounting to a violation of international law, to either (a) a breach of international law or (b) an unfriendly act, by another state"); L. Oppenheim, *International Law: A Treatise* (1912) at 36-37 ("The act which calls for retaliation is not an illegal act; on the contrary, it is an act that is within the competence of the doer").

33 Catherine Lotrionte, *Reconsidering the Consequences for State-Sponsored Hostile Cyber Operations Under International Law*, Cyber Defence Review (2018) at 92 ("An act of retorsion is a coercive, politically unfriendly, but lawful act, not involving any breach of international obligations owed to the target state, whether treaty-based or customary and thereby do not require any legal justification"); Lindsay Moir, *The Implementation and Enforcement of the Laws of Non-International Armed Conflict*, 3 J. Armed Conflict L. 163, 176 (1998) ("Retorsion is an unfriendly, even potentially damaging, act. Unlike reprisals, however, retorsion is perfectly valid under international law"); Lori Fisler Damrosch, *Enforcing International Law Through Non-Forcible Measures* (1997) at 54 (defining "retorsion" as "an unfriendly (but not otherwise illegal) act taken in response to an unfriendly or illegal act").

34 Troy Anderson, "*Fitting a Virtual Peg into a Round Hole: Why Existing International Law Fails to Govern Cyber Reprisals*, 34 Aris. J. Int'l & Compl L. 135, 142 (2016) ("Retorsion usually is diplomatic or economic in nature, rather than militaristic").

35 Ibid. at 147 ("Limiting a cyber operation to the confines of legality in order to allow it to qualify as a legal retorsion severely limits the power of the cyber operation").

36 Bruno Simma (ed), *The Charter of the United Nations: A Commentary*, (1994) at 104.

of another state.[37] This allows states to develop responses to adverse actions without first engaging in a legal analysis of whether the adversary's actions violated an international legal obligation.

Unlike countermeasures, retorsion is not necessarily confined to a particular goal. Some commentators have suggested that, as with countermeasures, it is often intended to persuade a state to cease its internationally wrongful acts,[38] but there is nothing in the Draft Articles or other authoritative commentary that would confine retorsion to mere persuasion. It also could include measures that blunt the harm of an adversary's actions, or prevent the adversary from exercising its capabilities. Indeed, the legal nature of retorsion means that it is not subject to the same limitations as other responses.[39] Relatedly, if a state engages in retorsion rather than countermeasures, it does not have a legal obligation to notify the other state.

Retorsion is not subject to the strict duration requirements of other responses. As described in the previous section, countermeasures must cease immediately once the other state has ceased its internationally wrongful acts. Oppenheim suggested in 1912 that because "retorsion is made use of only to compel a state to alter its discourteous, unfriendly, or unfair behaviour, all acts of retorsion ought at once to cease when such State changes its behaviour".[40] Even under this limited view – which is not supported by more recent authoritative sources – retorsion need not cease once the other state is *legally* compliant; indeed, a legal violation is not a precondition for retorsion. Oppenheim's suggestion, to the extent that it is followed, is a more pragmatic and political guideline: for instance, if State A's sanctions caused State B to stop attacking State A's election system, then it would be politically and diplomatically unwise for State A to continue the sanctions unless there was an indication of further malicious action by State B.

Nor does international law require retorsion to be proportionate to the malign actions of the adversary, as is required for countermeasures. *Brierly's Law of Nations* notes that "it is sometimes suggested that retaliation should be proportionate",[41] but it cites no binding or persuasive legal precedent that would suggest proportionality is a requirement for retorsion. As with the limit on duration, it may be that a disproportionate retorsion would raise political or diplomatic concerns, but proportionality is not a legal requirement of retorsion, which by definition is an independent legal act.

37 Andrew Clapham, *Brierly's Law of Nations* (7th ed. 2012) at 397 ("'Retorsion' is a measure of self-help taken in response to an illegal or unfriendly act, where the self-help measure itself is within the law").

38 Edward Kwakwa, *Belligerent Reprisals in the Law of Armed Conflict*, 27 Stanford J. Int'l L. 49, 51 (1990) ("a retorsion seeks to coerce another state to discontinue a vexatious or injurious – but legal – practice").

39 *Law of War Manual* (supra n 31) at 1110 ("Because retorsion, by definition, does not involve the resort to actions that would ordinarily be characterised as illegal, the stringent conditions that apply to reprisal do not apply to retorsion").

40 Oppenheim (supra n 31) at 38.

41 Clapham (supra n 37) at 397.

Unlike countermeasures, international law does not restrict nations from collaborating on retorsion. For instance, if a state has repeatedly acted maliciously in cyberspace with targets in multiple states, all of those states could collectively engage in sanctions or release a joint public statement condemning the bad actor.

In sum, retorsion is both narrow and broad. It is narrow in the sense that it only applies to actions that, standing independently, would not violate international law. It is broad because, if the acts qualify as retorsion, they are not subject to the same legal constraints as responses such as countermeasures. Retorsion is often overlooked in debates on international law, which tend to focus on countermeasures and self-defence. While those frameworks are vital to the discussion, we also must examine whether responses can constitute retorsion and are afforded more leeway.

4. ACTIONS THAT MAY CONSTITUTE RETORSION IN RESPONSE TO MALIGN CYBER OPERATIONS

Whether a response qualifies as retorsion depends entirely on whether the measure violates any international legal rules. Some actions such as intentionally causing damage to another state's computer systems do not constitute retorsion because they likely violate a legal obligation. But what *does* qualify as retorsion in the cyber realm? Legal scholarship often provides sanctions as the primary example of retorsion. Based on the definition of retorsion set out in this paper, sanctions in response to malign cyber actions clearly would qualify as retorsion. However, this paper posits that sanctions are only one form of retorsion, and policymakers should search more broadly for responses to cyber actions that are legal and therefore not subject to the same restrictions as countermeasures. This section categorises the types of responses that might fit into the broader concept of retorsion. To assess whether these actions qualify as retorsion, it is necessary to determine whether they violate sovereignty, prohibitions on the use of force or other legal norms.

A. Pressure via International Relations

The classic and most oft-cited examples of retorsion involve a target state using standard tools of international relations to pressure an adversary to stop its illegal or unfriendly acts. Such examples include "severance of diplomatic relations and the expulsion or restrictive control of aliens, as well as various economic and travel restrictions".[42] So too is a US law requiring suspension of foreign aid "to any country nationalising American property without proper compensation",[43] and the April 2015 executive order that allows sanctions for, among other things, "harming, or otherwise significantly compromising the provision of services by, a computer or network of

42 Malcom N. Shaw, *International Law* (8th ed. 2017) at 859.

computers that support one or more entities in a critical infrastructure sector".[44] In 2019, the US imposed sanctions on North Korean hackers accused of a number of cyber operations, including the 2014 hack of Sony Pictures.[45] Likewise, in December 2016, the US expelled 35 Russian diplomats in response to Russian interference in the 2016 US elections.[46]

A target state could also publicly shame a nation that has attacked its systems. For instance, countries are increasingly securing indictments in their domestic courts against foreign hackers.[47] In many cases, the countries issuing the indictments do not have extradition arrangements with the states where the hackers are located, so it is unlikely that the hackers will ever stand trial. However, the indictments play an important role in publicly "naming and shaming" both the individual cyber operators and, in many cases, the governments that employ them.

These responses are all, to varying degrees, unfriendly; yet they do not violate any international legal principles and therefore clearly qualify as retorsion. They do, however, face a number of political constraints. For example, if State A mistakenly attributes a DDOS attack to State B and implements sanctions, it risks significant diplomatic pushback from State B and other states. However, such a short-sighted act would not violate international law and is not subject to the same limits as countermeasures.

B. Accessing Information on the Adversary's Systems

A state that has been targeted by another state's hostile cyber acts may seek to access that state's systems to gather information about the adversary's operations. To the extent that this constitutes legal peacetime espionage, it should fall under the broad umbrella of retorsion. The general rule is that peacetime cyber espionage is not illegal per se.[48] Of course, a state still could violate another state's sovereignty if, for instance, an act of espionage causes damage to data or computer systems.[49] Moreover, although the prevailing view, as stated in the *Tallinn Manual*, is that peacetime cyber espionage

43 Ibid.

44 Executive Order 13694 (April 1, 2015).

45 Carol Morello & Ellen Nakashima, *US Imposes Sanctions on North Korean Hackers Accused in Sony Attack, Dozens of Other Incidents*, Wash. Post (Sept. 13, 2019), available at https://www.washingtonpost.com/national-security/us-sanctions-north-korean-hackers-accused-in-sony-attack-dozens-of-other-incidents/2019/09/13/ac6b0070-d633-11e9-9610-fb56c5522e1c_story.html.

46 David E. Sanger, *Obama Strikes Back at Russia for Election Hacking*, NY Times (Dec. 29, 2016), available at https://www.nytimes.com/2016/12/29/us/politics/russia-election-hacking-sanctions.html.

47 Alfred Ng, 'Justice Department Charges North Korean Over WannaCry, Sony Hack', CNET (Sept. 6, 2018), available at https://www.cnet.com/news/justice-department-charges-north-korean-hacker-linked-to-wannacry-2014-sony-hack/.

48 *Law of War Manual* at 1016 ("Generally, to the extent that cyber operations resemble traditional intelligence and counter-intelligence activities, such as unauthorised intrusions into computer networks solely to acquire information, then such cyber operations would likely be treated similarly under international law").

49 *Tallinn Manual* at 170 ("For instance, if organs of one State, in order to extract data, hack into the cyber infrastructure located in another State in a manner that results in a loss of functionality, the cyber espionage operation violates, in the view of the Experts, the sovereignty of the latter").

per se is legal, some governments have pushed back against this rule and argued that in some cases espionage may be an infringement of sovereignty.[50]

The current majority view, however, is that unless cyber espionage results in damage, it does not violate international law. To the extent that this remains the general rule, conducting espionage operations on an adversary's systems may constitute retorsion. For instance, assume that State A has repeatedly attempted to spread false information to interfere in the elections of State B. State B may attempt to access data on State A's computers that provides insight into this propaganda campaign. Assuming that this stays within the boundaries of legal peacetime cyber espionage, State B need not attempt to classify its action as a countermeasure, as it would constitute retorsion. Such characterisation is particularly useful in this scenario, as there is considerable debate as to whether such election interference constitutes a breach of international legal obligations. To the extent that State B's hacking operations can be characterised as retorsion, it need not concern itself with whether State A's campaign was legal.

The limited commentary about retorsion does not include espionage among the examples of retorsion, as the commentary typically focuses on responses such as sanctions and expulsion of diplomats. However, if we are to view retorsion as any unfriendly act that complies with international law, many forms of espionage would also fall within the definition of retorsion.

C. Conducting Cyber Operations on One's Own Network (Honeypots and Sinkholes)

The adversary's systems are not the only potential source of information about their capabilities and plans. A target state could learn about the adversary by observing their actions on the target state's own systems. Such operations are even more likely to qualify as retorsion than the espionage described above. Unlike the operations that take place on the adversary's systems, such local observations do not even come close to raising any questions of territorial sovereignty violations.

What if the target state took steps to entice the adversary to be present on its network, allowing the target state to observe the adversary's actions? For instance, imagine that State A has not only been launching propaganda to influence State B's elections, but also attempting to access and delete voting data from State B's elections systems. State B might use a honeypot[51] to lure State A to a particular State B server, and then observe State A's actions and gather information about its techniques. Honeypots

50 Russel Buchan, *The International Legal Regulation of State-Sponsored Cyber Espionage*,in International Cyber Norms: Legal, Policy & Industry Perspectives, Anna-Maria Osula, Henry Rõigas (eds.) (2016) at 71 ("There is state practice to suggest that where a state considers itself to have been the victim of cyber espionage it regards such behaviour as falling foul of the principle of territorial sovereignty").

51 Paul Rosenzweig, *International Law and Private Actor Active Cyber Defensive Measures*, 50 Stan. J. Int'l L. 103, 106 n8 (2014). ("As the name implies, honeypots are intended to attract hackers by purporting to be worthwhile subjects of attack. One might, for example, give a document honeypot the Microsoft Word name 'Plans for Countering Hackers.Docx' and expect it to be the subject of an attack").

can be quite useful both for obtaining information and distracting attackers on false systems.[52] Governments use two primary types of honeypots: production honeypots detect imminent threats and are easier to deploy, while research honeypots gather information about emerging tactics of adversaries.[53] Another tool, the sinkhole, diverts harmful traffic, such as a botnet, to prevent harm.[54]

This use of honeypots and sinkholes should qualify as retorsion because they do not involve the infringement of sovereignty or any other legal obligation to State A; indeed, the act takes place entirely on the systems of State B, once State A has accessed the system. Critics of this approach might argue that the use of such tools to deceive another state is more aggressive than traditional espionage. While this may be true, there is little support for a claim that such deception – occurring entirely on State B's systems as a result of State A's intentionally malicious actions – would violate international law. If, however, State B were to use data collected via the honeypot to cause damage to State A's systems, the act would probably no longer qualify as retorsion. State A could claim that State B caused damage by unnecessarily consuming State A's resources with a sinkhole, but that argument would not be likely to prevail because the distraction merely prevented State A from malicious acts against State B.

What if a state were to install malware in data exfiltrated from its network? Whether such an act would constitute retorsion would depend on the effects of the malware. If the malware merely allowed the target state to observe data on the adversary's network, such an action probably would constitute retorsion, as the impacts would be no different from other forms of espionage. However, if the malware caused damage to the adversary's systems, the act might not be classifiable as retorsion because it might violate the adversary's sovereignty, in which case it would need another justification such as countermeasures. The exact boundaries as to when honeypots constitute an internationally wrongful act are subject to significant debate.[55]

D. Influencing Adversaries

A state that has been targeted by state-sponsored hackers may seek to send a message to those hackers to discourage them from engaging in further such acts. Influence is one of the three primary operational components of the US Defense Department's

52 Ian Walden & Anne Flanagan, *Honeypots: A Sticky Legal Landscape*, 29 Rutgers Computer & Tech. L.J. 317, 319 (2003) ("It can serve as a decoy to deflect the hacker from breaking into the real system, as a research tool for systems administrators merely to observe and learn how hackers operate and about weaknesses in their systems, or as a tool to monitor and document evidence for criminal prosecution").

53 Josh Fruhlinger, *What is a Honeypot? A Trap for Catching Hackers in the Act*, CSO (April 1, 2019).

54 Lily Hay Newman, *Hacker Lexicon: What Is Sinkholing?* Wired (Jan. 2, 2010), available at https://www.wired.com/story/what-is-sinkholing/.

55 David Wallace & Mark Visger, *The Use of Weaponised 'Honeypots' Under the Customary International Law of State Responsibility*, Cyber Defence Review (Summer 2018) at 38 ("Moreover, is it not reasonable for a State defending its cyber infrastructure to take measures, like using honeypots, to protect itself against such intrusions and, quite frankly, deter others? Is it wrong for a State to use a dynamic, penalty-based form of deterrence? The law, as it is currently structured, does not address these questions").

operational concept of "Defend Forward".[56] For instance, in October 2018, the *New York Times* reported that US Cyber Command sent messages to Russian operatives who disseminated propaganda during US elections "telling them that American operatives have identified them and are tracking their work".[57]

Retorsion would be a particularly attractive classification for such an operation, as it would avoid the need to determine whether the Russian election propaganda constituted a breach of international law that justified countermeasures. Of course, the potential barrier to the retorsion classification would be a claim that the US messaging violated international law. As applied to the public reports of the Cyber Command operation, such an argument against retorsion is unlikely to succeed. The *Times* quoted senior defence officials anonymously stating that "they were not directly threatening the operatives",[58] so the operation likely does not raise any concerns about violating international humanitarian law.

A warning accompanied by a specific threat of physical injury to the hackers could violate sovereignty, prohibitions on threats of use of force and even international humanitarian law, but the target state could claim retorsion for a narrowly tailored message that simply makes the adversary aware the target is watching their actions. Such an action certainly is unfriendly, but it does not violate international legal obligations.

E. Establishing a Position on the Adversary's Network

If a state has been the target of malign cyber operations, it may seek to establish a position on the adversary's systems. Positioning, like influence, is a component of the US Defend Forward operational concept.[59] Such positioning serves two primary purposes. First, it might send a message to the adversary that further actions could have consequences. Robert Chesney has described such a move as a "hold at risk" operation, with the goal "of establishing access to a potential adversary's system is to bolster one's deterrence posture by making clear to the adversary that you are capable, as a practical matter, of overcoming their defences and harming something they value".[60] Second, establishing a position allows the target state to respond more

56 Jeff Kosseff, *The Contours of 'Defend Forward' Under International Law*, paper for the 2019 11th International Conference on Cyber Conflict (2019) at 5 ("The Defend Forward concept also encourages stability by disabusing adversaries of the idea that they can operate with impunity in cyberspace and signals US commitment to confront hostile activities and impose cumulative costs for ongoing malicious actions") (internal quotation marks and citations omitted).

57 Julian E. Barnes, *US Begins First Cyberoperation Against Russia Aimed at Protecting Elections*, NY Times (Oct. 23, 2018) available at https://www.nytimes.com/2018/10/23/us/politics/russian-hacking-usa-cyber-command.html.

58 Ibid.

59 Kosseff (supra n 56) at 5 "'Perhaps the biggest shift in US cyber operations under Defend Forward is Cyber Command's recognition of the need for a forward cyber posture that can be leveraged to persistently degrade the effectiveness of adversary capabilities and blunt their actions and operations before they reach US networks") (internal quotation marks and citation omitted).

60 Robert Chesney, *The 2018 DOD Cyber Strategy: Understanding 'Defense Forward' in Light of the NDAA and PPD-20 Changes*, Lawfare (Sept. 25, 2018), available at https://www.lawfareblog.com/2018-dod-cyber-strategy-understanding-defense-forward-light-ndaa-and-ppd-20-changes.

quickly to any further harmful cyber actions by the adversary, perhaps allowing it to disable the source of the malign actions. Chesney refers to this as a "preparation of the battlefield" operation.[61]

There is a strong argument that merely accessing the adversary's systems – either to hold at risk or to prepare the battlefield – does not constitute a wrongful act under international law and therefore can be categorised as retorsion. Even if one were to recognise a standalone sovereignty obligation that is separate from other rules such as non-intervention, it is far from certain that mere access would violate that obligation. To be sure, if the target state were to leverage that access, such as by causing harm to the adversary's system, such an action might raise sovereignty concerns and not be justifiable as retorsion. Accordingly, it is important to separate the legal analysis of establishing a position on a network from the analysis of using that position.

F. Slowing Down the Adversary

Cyber operations may also attempt to impede the progress of an adversary who has conducted malign cyber operations. While it might be possible to classify such operations as countermeasures, there is at least a reasonable chance that the actions do not violate international legal obligations and therefore constitute retorsion. Consider Operation Glowing Symphony, a 2016 operation in which US Cyber Command accessed ISIS media systems, "deleted files, closed accounts, changed passwords", and "began moving through the ISIS networks they had mapped for months like a raid team clearing a house".[62] The operation resulted in ISIS media operatives being locked out of their accounts, having slow connections and other glitches.[63] Would such actions be permissible if conducted against a nation-state? Even under the expansive view of territorial sovereignty, it is at least debatable whether such inconveniences amount to a violation of international law. Under the views articulated in the 2017 US Department of Defense internal memorandum, cyber operations are only constrained by the prohibitions on the use of force and on intervention, and therefore there is an even stronger argument that Operation Glowing Symphony complied with international legal obligations. Indeed, some commentators have speculated that the release of the Defense Department memo soon after disclosure of Operation Glowing Symphony "raises the possibility it was produced to instruct DoD components of the legal analysis that supported the operation".[64] Such slow-down operations are unfriendly, but absent more significant harms there is at least a reasonable argument that the operations are retorsion.

61 Ibid.

62 Dina Temple-Raston, How the US Cracked Into One of the Most Secretive Terrorist Organisations, NPR (Sept. 26, 2019), available at https://choice.npr.org/index.html?origin=https://www.npr.org/2019/09/26/764790682/how-the-u-s-cracked-into-one-of-the-most-secretive-terrorist-organizations?t=1588800161071.

63 Ibid.

64 Watts & Richard (supra n 15) at 862.

5. CONCLUSION

This paper has sought to better define retorsion in an effort to provide states with more certainty about their options if countermeasures or self-defence are impractical or unavailable. Retorsion is typically associated only with international affairs such as sanctions and public denunciation. While those are critical examples of retorsion, other responses should fall under the same umbrella. All legal responses should be available for states to impede and discourage malign cyber actions. Before proceeding to an analysis of how to qualify a response as a countermeasure, a state should first determine whether it can justify its response as retorsion.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Collective Countermeasures in Cyberspace – *Lex Lata*, Progressive Development or a Bad Idea?

Przemysław Roguski
Lecturer
Chair for Public International Law
Jagiellonian University
Kraków, Poland
przemyslaw.roguski@uj.edu.pl

Abstract: This paper analyses whether international law permits collective countermeasures against states responsible for cyberattacks. In her opening address at CyCon 2019, Estonia's President Kersti Kaljulaid presented Estonia's view that 'States which are not directly injured may apply countermeasures to support the state directly affected by the malicious cyber operation'. This view was rejected by France in its declaration of 9 September 2019 on how international law applies to cyber operations. Discussing the International Law Commission's treatment of the legality of third-party countermeasures in its Articles on State Responsibility, the paper finds that the question was ultimately left open, given the unsettled status of customary law at that time. However, the Articles are formulated in such a way as to allow the application of lawful measures by not directly injured States, thus leaving room for developments in international law. Based on recent scholarship and examples of State practice, the paper finds that international law has indeed evolved since 2001 to permit collective countermeasures, but only insofar as third-party countermeasures against violations of collective obligations are concerned. In consequence, collective action by non-injured States against cyberattacks violating the sovereignty of a State or constituting an intervention in its internal affairs are not permitted under international law as it stands today. Lastly, the paper discusses whether international law may

recognise cyber-specific collective obligations and finds that the obligation to protect the 'public core of the internet' may be a good candidate for such a norm.

Keywords: *collective countermeasures, state responsibility, erga omnes, community interest, public core of the internet*

1. INTRODUCTION

Imagine the following scenario: State A suffers a series of cyberattacks against its critical infrastructure (electricity supply stations, public transportation, etc.). The attacks are attributed to State B, a much larger, technologically advanced and economically more powerful State. State A lacks the technical capacity to actively defend against the cyberattacks and fears that 'offline' countermeasures would not be effective if undertaken alone. Luckily, State A is part of a larger union of like-minded States and asks its partners for assistance in stopping the cyberattacks by adopting collective countermeasures against State B. After all, '[a]llies matter also in cyberspace'.[1]

This or a similar scenario might have motivated Estonia's President Kersti Kaljulaid to further the position that 'States which are not directly injured may apply countermeasures to support the state directly affected by the malicious cyber operation'.[2] While initial reactions from academia were positive,[3] other States' reactions were much more muted. The Dutch Minister of Foreign Affairs' letter of 5 July 2019 to the President of the House of Representatives setting out the Dutch government's view of the international legal order in cyberspace does not mention the possibility of collective countermeasures at all,[4] while the French document on international law applicable to cyber operations, perhaps the most elaborate reflection on the applicability of international law in cyberspace today, rules out the possibility

1 Kersti Kaljulaid, President of the Republic at the opening of CyCon 2019, Speech in Tallinn on 29 May 2019, https://www.president.ee/en/official-duties/speeches/15241-president-of-the-republic-at-the-opening-of-cycon-2019/index.html [19.04.2020].

2 Ibid.

3 See, e.g., Michael Schmitt, 'Estonia Speaks out on Key Rules for Cyberspace' (*Just Security*, 10 June 2019) https://www.justsecurity.org/64490/estonia-speaks-out-on-key-rules-for-cyberspace/ [19.04.2020].

4 Dutch Ministry of Foreign Affairs, *Letter to the parliament on the international legal order in cyberspace*, https://www.government.nl/ministries/ministry-of-foreign-affairs/documents/parliamentary-documents/2019/09/26/letter-to-the-parliament-on-the-international-legal-order-in-cyberspace [19.04.2020].

of France taking part in collective countermeasures based on the view that under current international law such measures may only be taken by the victim State.[5]

This paper aims to examine whether current international law today permits the application of countermeasures not only by the State-victim of a cyberattack, but also by non-injured States as a 'solidarity measure' to induce the responsible State to abide by international law. The paper is in four parts. First, it will explore the drafting history and current text of the International Law Commission's Articles on State Responsibility[6] to establish whether they allow for the implementation of collective countermeasures. Next, it will examine current State practice with respect to collective countermeasures and cyber-specific collective action to inquire how these findings apply to cyberspace. Third, it will examine whether there are any collective obligations[7] in cyberspace and lastly, it will offer a conclusion and an outlook concerning the question of whether a progressive development of international law to include collective countermeasures would be a good idea.

2. COLLECTIVE COUNTERMEASURES AND THE ILC'S ARTICLES ON STATE RESPONSIBILITY

A. Standing to Invoke the International Responsibility of a State

It is a fundamental principle of international law, indeed of law itself, that any breach of an obligation gives rise to a responsibility on the subject found in breach of that obligation.[8] In most national legal systems, the competence to invoke this responsibility lies with the natural or legal person to whom the obligation was owed or, if the obligation is owed to society or society has a particular interest in ensuring respect for certain obligations (as in criminal or administrative law, for example), with the State. However, the enforcement of responsibility is limited solely to the State due to its internal sovereignty and exclusive competence to create and enforce the legal system applicable in that State. This is different in international law. Because the concept of State responsibility is a necessary corollary of State sovereignty,[9] it is precisely the sovereign equality of States which puts limits on the competence to invoke and enforce the international responsibility of another State. Thus, in the traditional 'Westphalian' system the international community did not possess

5 French Ministry of the Armies, *International Law Applied to Operations in Cyberspace*, https://www.defense.gouv.fr/content/download/567648/9770527/file/international+law+applied+to+operations+in+cyberspace.pdf [19.04.2020].

6 International Law Commission, *Draft articles on Responsibility of States for Internationally Wrongful Acts, with commentaries*, Supplement No. 10 (A/56/10), ch.IV.E.1 (ARSIWA Commentaries).

7 Which are to be understood as obligations applicable between a group of States and established in some collective interest, ARSIWA Commentaries, Article 48 para 7.

8 Cf. F*actory at Chorzów*, Judgment, 1928, PCIJ, Series A, No. 17, 4, 29; Dionisio Anzilotti, *Cours de Droit International* (Recueil Sirey 1929) 467.

9 Alain Pellet, 'The Definition of Responsibility in International Law' in James Crawford and others (eds), *The Law of International Responsibility* (Oxford University Press 2010) 4.

centralised enforcement structures and the invocation of responsibility was a bilateral matter.[10] In other words, only States which were directly injured by the violation of an international obligation have standing to invoke the international responsibility of the violating State.

To induce or coerce the violating State to abide by its international obligations towards the injured State, the latter employed a series of measures ranging from actions which are unfriendly or hostile (retorsions) to actions which normally are unlawful under international law, but are permitted because of the previous violation of an international obligation (reprisals).[11] While belligerent reprisals are prohibited under Art. 2(4) of the UN Charter and the authority to enforce certain international obligations which are important for the preservation of international peace and security – such as the prohibition of the use of force – rests with the UN Security Council, the enforcement of bilateral obligations largely remains within the sphere of bilateral relations as a mechanism of 'private justice'.[12]

B. Discussion within the International Law Commission

Working on a mandate from the UN General Assembly to codify the principles of international law governing State responsibility,[13] when the International Law Commission (ILC) considered the question of countermeasures,[14] it also addressed whether such measures can only be taken bilaterally by the injured State against the responsible State or whether they may also be taken by other States.[15] The 1996 draft of the Articles on State Responsibility was based on the bilateral model of responsibility where even the violation of multilateral obligations could lead only to bilateral enforcement between the injured and the responsible State.[16] This has been found unsatisfactory by the ILC and, in particular, Special Rapporteur James Crawford, who believed that 'countermeasures are no longer limited to breaches of bilateral obligations, or to responses taken by the State most directly injured',[17] but may be permissible against breaches of obligations *erga omnes*, i.e. actions deemed an offence against all members of the international community.[18]

10 Ibid 6–7.

11 Matthias Ruffert, 'Reprisals', *Max Planck Encyclopaedia of Public International Law* (2015) para 2; Shane Darcy, 'Retaliation and Reprisal' in Marc Weller (ed), *The Oxford Handbook of the Use of Force in International Law* (Oxford University Press 2015) 880.

12 Denis Alland, 'Countermeasures of General Interest' (2002) 13 European Journal of International Law 1221, 1226.

13 UN General Assembly Resolution 799 of 7 December 1953, UN Doc. A/Res/799(VIII).

14 Which came to signify 'unilateral' or 'horizontal' reactions of one or more states to an internationally wrongful act, to the exclusion of self-defence and retorsion', Gaetano Arangio-Ruiz (Special Rapporteur), *Third Report on State Responsibility* (1991), UN Doc. A/CN.4/440 and Add.1, para 27.

15 For a study of the work of the International Law Commission on countermeasures see Alland (n 12); Martti Koskenniemi, 'Solidarity Measures: State Responsibility as a New International Order?' (2002) 72 British Yearbook of International Law 337.

16 Martin Dawidowicz, *Third-Party Countermeasures in International Law* (Cambridge University Press 2017) 89.

17 James Crawford (Special Rapporteur), *Second Report on State responsibility* (1999), UN Doc. A/CN.4/498 and Add.1-4, para 247.

18 Ibid.

In his third report,[19] Crawford understood the term 'collective countermeasures' to mean the right to react – in the public interest – against breaches of collective obligations to which the reacting States are parties, even though they were not individually injured by the breach.[20] It is important to stress that these collective countermeasures only refer to reactions taken by one State or by a group of States each acting in its individual capacity, and not institutional reactions within the framework of international organisations such as the United Nations.[21] After an examination of State practice, the Special Rapporteur concluded that there were a 'considerable number of instances' where non-injured States 'have taken measures against a target State in response to prior violations of collective obligations by that State'.[22] Examples included the trade embargo imposed by the European Community, Australia, Canada and New Zealand against Argentina after it invaded the Falkland Islands, or those against Iraq after its invasion of Kuwait.[23] However, he admitted that practice does not allow 'clear conclusions to be drawn as to the existence of a right of States to resort to countermeasures in the absence of injury'.[24] Nevertheless, Crawford saw support for the view that a State which was injured by a breach of a multilateral obligation 'should not be left alone to seek redress for the breach'.[25] Crawford's proposals were taken up by the Drafting Committee, which included them in Draft Article 54 [2000] to the effect that 'Any State entitled […] to invoke the responsibility of a State may take countermeasures at the request and on behalf of any State injured by the breach'.[26]

However, in the ensuing debate in the ILC, the views on collective countermeasures were split. Supporters claimed that the main purpose of collective countermeasures was to provide a viable alternative to the use of force and was the essential consequence of serious breaches of community norms without which States would be powerless to enforce these norms.[27] Opponents argued twofold: first, that the existing State practice did not support the conclusion that international law allows imposition of countermeasures by non-injured States and, second, that serious breaches of obligations owed to the international community as a whole were in principle a matter for the UN Security Council.[28] Views were similarly split in the debate in the Sixth Committee of the UN General Assembly, with some States supporting Draft

19 James Crawford (Special Rapporteur), *Third Report on State responsibility* (2000), UN Doc. A/CN.4/507 and Add.1-4, paras 386-405.
20 Ibid. para 386.
21 Ibid. para 387.
22 Ibid. para 395.
23 Ibid. para 391.
24 Ibid. para 397.
25 Ibid. para 401.
26 ILC Report (2000), UN Doc. A/55/10, 70.
27 See e.g. International Law Commission, *Summary records of the meetings of the fifty-third session 23 April – 1 June and 2 July – 10 August 2001*, Yearbook of the International Law Commission 2001, Vol I, 41, para 49 (Mr. Pellet).
28 Ibid. 35, para 2 (Mr. Brownlie, calling collective countermeasures 'neither *lex lata* nor *lex ferenda* [but] *lex horrenda*'); Ibid. 34 (Mr. Sepúlveda-Amor).

Article 54 [2000] and others voicing concerns about the potential abuse of collective countermeasures by powerful States[29] and potential conflict with the competences of the UN Security Council.[30] In the end, due to the difficult problems raised by the concept of collective countermeasures, some States proposed to accommodate the differing views by replacing Draft Article 54 [2000] with a savings clause.[31]

C. Final Draft of the Articles on State Responsibility

The ILC ultimately decided not to take a position on collective countermeasures, admitting that 'there appears to be no clearly recognised entitlement of [non-injured States] to take countermeasures in the collective interest'.[32] In consequence, under Art. 48 ARSIWA, States other than the injured State may invoke the international responsibility of another State if it breaches a collective obligation, i.e. if the obligation breached is owed to a group of States and established for the protection of a collective interest of the group (Article 48(1)(a))[33] or if it breaches an obligation owed to the international community as a whole (Article 48(1)(b)).[34] However, invoking responsibility under this provision is limited to requesting of the responsible State cessation, non-repetition or performance or a combination thereof (Article 48(2)), all of which stops short of permitting any enforcement action by the non-injured State. Additionally, a savings clause was inserted into Art. 54 ARSIWA to the effect that nothing in the chapter on countermeasures prejudices the right of a State entitled under Art. 48 to take 'lawful measures' to ensure cessation and reparation.

Two major conclusions can be drawn from the analysis so far. First, that the Articles on State Responsibility in their current form do not endorse, but neither do they preclude the imposition of countermeasures by groups of States other than the injured State. Such collective countermeasures would, therefore, be lawful if it were established that there is sufficient State practice and *opinio iuris* to support the existence of an international customary rule allowing for collective countermeasures. Second, however, under Art. 48 ARSIWA, non-injured States would only have standing to invoke the responsibility of another State if the obligation breached is either owed to a group of States and established for the protection of collective interest (so-called *erga omnes partes* obligations)[35] or to the international community as a whole (so-called *erga omnes* obligations).[36] The next steps of the analysis will, therefore, examine whether international law has evolved to include a customary norm allowing for

29 UN Doc. A/C.6/55/SR.18, 11, para 59-62 (Cuba); UN Doc. A/C.6/55/SR.18, 9, para 51 (Russia); UN Doc. A/C.6/55/SR.15, 5–6, paras 29, 31 (India).

30 UN Doc. A/C.6/55/SR.22, 8, para 52 (Libya); UN Doc. A/C.6/55/SR.15, 3, para 17 (Iran); UN Doc. C.6/56/SR.16, 7, para 40 (Colombia); UN Doc. A/C.6/55/SR.24, 11, para 64 (Cameroon).

31 UN General Assembly Sixth Committee, *Summary record of the 14th meeting*, UN Doc. A/C.6/55/SR.14, 7 para 32 (United Kingdom).

32 ARSIWA Commentaries, Art. 54 para 6.

33 For instance, regional human rights treaties or nuclear-free-zones, see ARSIWA Commentaries, Art. 48 para 7.

34 For instance, the prohibition of aggression or genocide, see ARSIWA Commentaries, Art. 48 para 9.

35 ARSIWA Commentary, Art. 48 para 6.

36 ARSIWA Commentary, Art. 48 para 8.

collective countermeasures and whether cyberattacks may violate the abovementioned types of collective obligations.

3. COLLECTIVE COUNTERMEASURES IN INTERNATIONAL PRACTICE

A. Collective Countermeasures in post-2000 State Practice

In recent years, two major studies by Katselli Proukaki[37] and Dawidowicz,[38] and several shorter analyses[39] have examined whether collective or third-party countermeasures are permissible under customary international law. Both Dawidowicz and Katselli Proukaki have given extensive examples of measures instituted by States not directly injured by the violation of community norms against the responsible State, which may be characterised as countermeasures, including many which were not considered by the ILC.[40] Post-2001 (i.e. after the ILC Articles on State Responsibility were adopted), they list, for instance, collective action by the European Union and 26 other States against Burma (as it then was),[41] by various Western and Arab States against Syria[42] and most recently by Western States against Russia.[43]

The actions against Syria followed President Bashar al-Assad's violent suppression of peaceful protests in 2011 and the subsequent civil war, in which the Syrian regime committed countless atrocities and breaches of human rights and IHL norms. Sanctions against Syrian officials and the Syrian State have been imposed by the EU, 10 other European States and by the US.[44] These included freezing the assets of the Central Bank of Syria, which are otherwise immune from seizure under customary rules of State immunity.[45] The League of Arab States and its successor the Arab League have also imposed sanctions on Syria, ranging from the exclusion from participation in League meetings (for which there is no clear foundation in the applicable treaty) to freezing Syrian government assets and a ban on civil aviation.[46] Based on the measures imposed, which included the violation of treaty obligations and other applicable

37 Elena Katselli Proukaki, *The Problem of Enforcement in International Law* (Routledge 2010).
38 Dawidowicz (n 16).
39 Koskenniemi (n 15); N Jansen Calamita, 'Sanctions, Countermeasures, and the Iranian Nuclear Issue' (2009) 42 Vanderbilt Journal of Transnational Law 1393; Martin Dawidowicz, 'Public Law Enforcement without Public Law Safeguards? An Analysis of State Practice on Third-Party Countermeasures and Their Relationship to the UN Security Council' (2010) 77 British Yearbook of International Law 333; Carlo Focarelli, 'International Law and Third-Party Countermeasures in the Age of Global Instant Communication' (2016) 29 Questions of International Law 17 <http://www.qil-qdi.org/international-law-third-party-countermeasures-age-global-instant-communication/>.
40 Katselli Proukaki (n 37) 110–201; Dawidowicz (n 16) 112–238.
41 Dawidowicz (n 16) 196; Katselli Proukaki (n 37) 191.
42 Dawidowicz (n 16) 220–232.
43 Ibid. 231–238.
44 Ibid. 223-224.
45 Ibid. 222.
46 Ibid. 225.

norms of international law, the only reasonable justification for these actions is their character as third-party countermeasures.

The most recent example of the imposition of collective countermeasures is the case of the Russian annexation of Crimea. The facts are well known, but it is useful to briefly recall that by sending troops into Crimea and conducting an illegal referendum which ended in the annexation of the region, Russia committed acts which a majority of commentators consider to constitute an act of aggression and a violation of the right to self-determination of the Ukrainian people.[47] As both norms have *erga omnes* character, States other than Ukraine may also impose countermeasures to induce Russia to respect Ukrainian sovereignty and self-determination. Consequently, both the US and EU have imposed restrictive measures against certain Russian citizens involved in the takeover of Crimea and unilateral sanctions against Russia's defence, energy and financial sectors.[48] As financial transactions are covered by GATS, the EU measures have to be regarded as violations of an international obligation, but their wrongfulness is precluded due to their character as countermeasures.[49] In consequence, the measures adopted against Russia by certain States may serve as other examples of third-party countermeasures against violations of *erga omnes* obligations.[50]

These examples, and others analysed by Dawidowicz and Katselli Proukaki, demonstrate that there is widespread post-2001 State practice which seems to support the conclusion that customary international law does permit the imposition of collective countermeasures against violations of obligations *erga omnes (partes)*. Importantly, the Syrian example shows that not only Western States, but also the wider international community use sanctions as instruments to induce compliance with the most important community obligations. The present author would agree that there are indeed many examples of impositions of restrictive measures and sanctions by States, but it has to be noted that most of those examples refer to actions taken by Western States, which might suggest that State practice is predominately 'Western' and thus not sufficiently universal to create a norm of customary international law. However, examples of non-Western collective countermeasures – such as in Syria – also exist. Unfortunately, less frequent are statements of *opinio iuris* by the acting States which would allow us to understand the legal basis for particular collective actions. A few such statements do exist and, as Dawidowicz argues, normative intent can also be deduced from consistent practice.[51] In consequence, the argument can

47 See e.g. Veronika Bílková, 'The Use of Force by the Russian Federation in Crimea' (2015) 75 Zeitschrift für ausländisches öffentliches Recht und Völkerrecht 27; Christian Marxsen, 'The Crimea Crisis: An International Law Perspective' (2014) 74 Zeitschrift für ausländisches öffentliches Recht und Völkerrecht 367.

48 Martin Dawidowicz, 'Third-Party Countermeasures: A Progressive Development of International Law? - QIL QDI' (2016) 29 Questions of International Law 3.

49 Ibid.

50 Maurizio Arcari, 'International Reactions to the Crimea Annexation under the Law of State Responsibility: 'Collective Countermeasures' and Beyond?' in Władysław Czapliński and others (eds), *The Case of Crimea's Annexation under International Law* (Scholar 2017) 228f.

51 Dawidowicz (n 16) 253–254.

be made that collective countermeasures against States committing breaches of *erga omnes (partes)* obligations are lawful under customary international law, and thus not precluded by the Articles on State Responsibility by virtue of the savings clause of Article 54 ARSIWA.

B. Statements on International Law in Cyberspace

Both the GGE Report of 2015[52] and individual States confirm the general applicability of the law of State responsibility to State actions in cyberspace. However, other than stating that 'States must meet their international obligations regarding internationally wrongful acts attributable to them under international law',[53] the GGE Report gives no guidance on how certain concepts of State responsibility apply. In addition, there was significant opposition from some States in the 2016–17 GGE against the inclusion in the report of more specific references to countermeasures which, in the end, was not adopted.[54] However, no State argued for a cyberspace-specific *lex specialis* of State responsibility.[55]

Until January 2020, only two States had addressed the question of collective countermeasures. One is Estonia, which argued for the permissibility of collective countermeasures, subject to proportionality and as a means of last resort, where diplomatic action is insufficient and no lawful recourse to the use of force exists.[56] It has to be noted that Estonia did not claim that collective countermeasures are already permissible under international law, but was 'furthering' the position that non-injured States may apply countermeasures. The other State which has presented its views on collective countermeasures is France, which rejected the applicability of collective countermeasures in cyberspace.[57] France argued that under current international law collective countermeasures are not authorised, 'which rules out the possibility of France taking such measures in response to an infringement of another State's rights'.[58] Given that no other States have declared their views on this matter thus far, no clear common position can be discerned and the issue remains contentious.

52 UN General Assembly, *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*, 22 July 2015, UN Doc. A/70/174 ['GGE Report 2015'].

53 GGE Report 2015, para 28(f).

54 Barrie Sander, 'Democracy under the Influence: Paradigms of State Responsibility for Cyber Influence Operations on Elections' (2019) 18 Chinese Journal of International Law 1, 30.

55 Michael N Schmitt and Liis Vihul, 'International Cyber Law Politicized: The UN GGE's Failure to Advance Cyber Norms' (*Just Security*, 30 June 2017) https://www.justsecurity.org/42768/international-cyber-law-politicized-gges-failure-advance-cyber-norms/ [19.04.2020].

56 Kersti Kaljulaid, *President of the Republic at the opening of CyCon 2019*, Speech in Tallinn on 29 May 2019, https://www.president.ee/en/official-duties/speeches/15241-president-of-the-republic-at-the-opening-of-cycon-2019/index.html [19.04.2020].

57 French Ministry of the Armies, *International Law Applied to Operations in Cyberspace*, https://www.defense.gouv.fr/content/download/567648/9770527/file/international+law+applied+to+operations+in+cyberspace.pdf [19.04.2020].

58 Ibid. 7.

C. Examples of Cyber-Specific Collective Action

In recent years, States have in certain instances started to coordinate their responses to cyberattacks. The most notable forms of cooperation include collective attribution and cyber restrictive measures, which may be employed by EU Member States against the perpetrators of cyberattacks. However, none of these examples of collective action in cyberspace can be qualified as collective countermeasures.

1) Collective Attributions

While attribution by individual States can take many forms including criminal indictments, economic sanctions, technical alerts or official statements,[59] collective attributions mostly take place as a series of coordinated statements or press releases by a number of States. For instance, in December 2017 the UK,[60] US,[61] Australia,[62] Canada,[63] New Zealand[64] and Japan[65] released coordinated statements attributing the WannaCry ransomware attack to North Korea. Similar coordinated attributions followed the NotPetya cyberattacks,[66] the Russian hacking attempt of the OPCW which was jointly denounced by the UK and the Netherlands[67] and most recently the Russian cyberattacks against Georgia in 2018.[68] It has to be noted that public attributions alone, if not followed by enforcement action, do not infringe a State's rights under international law because international law does not prohibit one State from commenting on another State's actions, as long as such comments do not amount to coercion in regard to the other State's internal affairs.[69] Since they do not infringe

59 Kristen E Eichensehr, 'The Law & Politics of Cyberattack Attribution' (2019) UCLA School of Law Public Law Research Paper No. 19–36 10, <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3453804> [12.01.2020].

60 U.K. Foreign & Commonwealth Office, *Foreign Office Minister condemns North Korean actor for WannaCry attacks* (Press release of 19 December 2017) https://www.gov.uk/government/news/foreign-office-minister-condemns-north-korean-actor-for-wannacry-attacks [19.04.2020].

61 The White House, *Press Briefing on the Attribution of the WannaCry Malware Attack to North Korea* (Press briefing of 19 December 2017) https://www.whitehouse.gov/briefings-statements/press-briefing-on-the-attribution-of-the-wannacry-malware-attack-to-north-korea-121917/ [19.04.2020].

62 Australian Minister of Foreign Affairs, Press statement of 20 December 2017, https://dfat.gov.au/international-relations/themes/cyber-affairs/Documents/australia-attributes-wannacry-ransomware-to-north-korea.pdf [19.04.2020].

63 Government of Canada, *CSE Statement on the Attribution of WannaCry Malware*, https://www.cse-cst.gc.ca/en/media/2017-12-19 [19.04.2020].

64 New Zealand National Cyber Security Centre, *New Zealand concerned at North Korean cyber activity*, 20 December 2017, https://www.ncsc.govt.nz/newsroom/new-zealand-concerned-at-north-korean-cyber-activity/ [19.04.2020].

65 Ministry of Foreign Affairs of Japan, *Press statement by press secretary Norio Maruyama of 20 December 2017*, https://www.mofa.go.jp/press/release/press4e_001850.html [19.04.2020].

66 Eichensehr (n 59) 17.

67 United Kingdom and Kingdom of the Netherlands, *Joint statement from Prime Minister May and Prime Minister Rutte*, Press release of 4 October 2018, https://www.gov.uk/government/news/joint-statement-from-prime-minister-may-and-prime-minister-rutte [19.04.2020].

68 UK Foreign & Commonwealth Office, *UK condemns Russia's GRU over Georgia cyber-attacks* (Press release of 20 February 2020) https://www.gov.uk/government/news/uk-condemns-russias-gru-over-georgia-cyber-attacks [19.04.2020]; US Department of State, *The United States Condemns Russian Cyber Attack Against the Country of Georgia* (Press statement of 20 February 2020) https://www.state.gov/the-united-states-condemns-russian-cyber-attack-against-the-country-of-georgia/ [19.04.2020].

69 *Case concerning military and paramilitary activities in and against Nicaragua* (Nicaragua v. United States of America), Judgment, 27 June 1986, ICJ Rep. 1986, 14, para 205.

another State's rights, these public attributions do not constitute internationally wrongful acts which would need to be justified under the doctrine of countermeasures. At most, they might qualify as retorsions, i.e. reactions which do not interfere with the target State's rights under international law.[70]

2) Cyber Restrictive Measures

On 17 May 2019, the Council of the European Union adopted its decision concerning restrictive measures against cyber-attacks threatening the Union or its Member States.[71] By virtue of this decision, the Union and the Member States may apply restrictive measures (i.e. sanctions) against natural or legal persons who are responsible for (attempted) cyber-attacks with (potentially) significant effect constituting an external threat to the Union or its Member States (Article 1). These sanctions include travel bans on natural persons (Article 4) and the freezing of assets of natural and legal persons (Article 5). As both the travel bans and asset freezes affect the rights of individuals and entities present on the territory of EU Member States, the regulations fall within their territorial jurisdiction and therefore the imposition of such restrictive measures does not normally violate obligations owed to other States. In cases where it might, for instance with respect to the immunities of a person affected by restrictive measures, the Council Decision provides a series of exceptions (Article 4(3)). Therefore, it has to be concluded that these restrictive measures, even if imposed collectively, cannot be regarded as countermeasures.

3) 'Persistent Engagement' and 'Defending Forward'

It is, however, possible that States conduct cyber operations which may be qualified as third-party countermeasures. Under the doctrine of persistent engagement, the US has begun to take a pro-active stance in cyberspace and to 'maintain a forward presence' there.[72] This includes working with allies in friendly and foreign networks to counter malicious cyber operations against them.[73] If such operations are conducted against the network of the perpetrator State of a cyberattack and in response to a prior violation of international law owed to the affected third State by the targeted State, scholarly opinion[74] and certain States[75] might qualify this as a violation of the target State's sovereignty, rendering the operation a (third-party) countermeasure if conducted in response to a prior violation of international law by the targeted State. However, at the

70 Thomas Giegerich, 'Retorsion', *Max Planck Encyclopaedia of Public International Law* (Oxford University Press 2011) para 1.

71 Council Decision (CFSP) 2019/797 of 17 May 2019 concerning restrictive measures against cyber-attacks threatening the Union or its Member States, ST/7299/2019/INIT, OJ L 129I, 17.5.2019, 13–19.

72 Mark Pomerlau, 'Two Years in, How Has a New Strategy Changed Cyber Operations?' (*Fifth Domain*, 11 November 2019) https://www.fifthdomain.com/dod/2019/11/11/two-years-in-how-has-a-new-strategy-changed-cyber-operations/ [19.04.2020].

73 Ibid.

74 Michael N Schmitt and Liis Vihul (eds), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press 2017) 17.

75 E.g. France, see French Ministry of the Armies, *International Law Applied to Operations in Cyberspace*, https://www.defense.gouv.fr/content/download/567648/9770527/file/international+law+applied+to+operations+in+cyberspace.pdf [19.04.2020], 6.

time of writing, the US has not admitted to having conducted cyber operations against foreign targets in defence of a third State's rights under international law. Therefore, the existence of State practice in this regard cannot be confirmed from open sources.

4. IN SEARCH OF COLLECTIVE OBLIGATIONS IN CYBERSPACE

Based on the preceding findings two conclusions must be drawn. First, there is sufficient practice to support the finding that after the adoption of the ILC Articles on State Responsibility, international law has evolved to accept the imposition of not only individual but also collective countermeasures. Second, however, collective countermeasures are only permissible against violations of *collective* obligations. It is, therefore, necessary to inquire whether cyberattacks may violate such collective obligations.

A. Do 'Typical' Cyber Operations Violate Collective Obligations?

International instruments do not provide a definitive list of collective obligations of States. The ARSIWA Commentaries clarify that collective obligations under Art. 48(1)(a) ARSIWA, sometimes also referred to as 'obligations *erga omnes partes*', must transcend the sphere of bilateral relations of the States parties to the treaty establishing that obligation. Such obligations must protect a collective interest over and above the individual interests of States.[76] Examples of community interests protected by international law may include the protection of common goods in international environmental law,[77] standards of protection for a group of people, especially within human rights law,[78] or international common spaces such as the moon or celestial bodies.[79] The International Court of Justice confirmed that, for instance, the Genocide Convention serves a common interest rather than individual interests of States.[80] Similarly, obligations under Art. 48(1)(b) ARSIWA, are owed to the international community as a whole and all States have a legal interest in their protection.[81] These obligations *erga omnes* include the prohibition of aggression and genocide, protection of basic rights of the human person, including protection from slavery and racial discrimination,[82] the right of peoples to self-determination[83] and fundamental rules of international humanitarian law.[84]

76 ARSIWA Commentaries Art. 48 para 7.
77 Ibid.; Isabel Feichtner, 'Community Interest', *Max Planck Encyclopaedia of Public International Law* (Oxford University Press 2007) para 15.
78 Ibid. para 19.
79 Ibid. para 24.
80 *Reservations to the Convention on the Prevention and Punishment of the Crime of Genocide*, Advisory opinion, [1951] ICJ Rep 15, 23.
81 ARSIWA Commentaries Art. 48 para 8; *Barcelona Traction, Light and Power Company Limited* (Belgium v Spain), Judgment, [1970] ICJ Rep 3, para 33.
82 Ibid. para 34.
83 *East Timor* (Portugal v Australia), Judgment, [1995] ICJ Rep 90, para 29.
84 *Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory*, Advisory Opinion, [2004] ICJ Rep 136, paras 155-159.

With this in mind, it seems quite clear that the vast majority of cyber operations do not violate collective obligations. It may be conceivable that certain types of cyberattacks during armed conflict would violate IHL rules such as the principles of proportionality or distinction. However, it is rather unlikely that peacetime cyber operations would breach environmental rules or the prohibitions against torture, slavery or genocide. State declarations on the applicability of international law to cyber operations typically discuss whether cyber operations may violate the prohibition on the use of force, the principle of non-intervention and territorial sovereignty.[85] None of these rules are established for the protection of community interests (possibly with the exception of Art. 2(4) UN Charter), as there is no community interest in the non-interference in the internal affairs or territorial sovereignty of a particular State; rather they protect individual rights of affected States. Thus, a breach of these norms may not be invoked by non-injured States to institute (collective) countermeasures against the responsible State.

B. The Obligation to Protect the 'Public Core of the Internet' as a Potential Cyber-specific Community Interest Norm

It is, however, possible that cyber-specific community interests exist. A potential cyber-specific norm serving the community interest of all States may be the obligation to protect the 'public core of the internet'.

1) The Concept of the Public Core of the Internet

The concept of the 'public core of the internet' was first introduced in a report written for the Netherlands Scientific Council for Government Policy by Dennis Broeders.[86] The report made the argument that certain parts of the internet – its main protocols and infrastructure, which are responsible for the interoperability of networks and the global availability of content, services and resources – constitute the internet's 'public core',[87] which is increasingly under threat of disruptive action by States.[88] However, given the importance of the internet in today's world, those parts of the internet which guarantee its universality, interoperability, accessibility, integrity, availability and confidentiality and therefore its functioning as a global system should be regarded as a global public good and protected from interference.[89]

85 See e.g. the French or Dutch declarations, French Ministry of the Armies, *International Law Applied to Operations in Cyberspace*, https://www.defense.gouv.fr/content/download/567648/9770527/file/international+law+applied+to+operations+in+cyberspace.pdf [19.04.2020], 6-8; Dutch Ministry of Foreign Affairs, *Letter of 5 July 2019 from the Minister of Foreign Affairs to the President of the House of Representatives on the international legal order in cyberspace*, https://www.government.nl/binaries/government/documents/parliamentary-documents/2019/09/26/letter-to-the-parliament-on-the-international-legal-order-in-cyberspace/International+Law+in+the+Cyberdomain+-+Netherlands.pdf [19.04.2020], 1-4.

86 Dennis Broeders, *The Public Core of the Internet* (Amsterdam University Press 2015).

87 Dennis Broeders, 'Aligning the International Protection of 'the Public Core of the Internet' with State Sovereignty and National Security' (2017) 2 Journal of Cyber Policy 366, 2.

88 Broeders (n 86) 10.

89 Ibid. 45.

The idea was taken up and further developed by the cyber policy community. In November 2017 the Global Commission on the Stability of Cyberspace (GCSC) issued a call to protect the public core of the internet, which stated that 'without prejudice to their rights and obligations, state and non-state actors should not conduct or knowingly allow activity that intentionally and substantially damages the general availability or integrity of the public core of the internet, and therefore the stability of cyberspace'.[90] The public core of the internet was also endorsed in the Paris Call for Trust and Security in Cyberspace of 12 November 2018,[91] which included a commitment to implementing cooperative measures to '[p]revent activity that intentionally and substantially damages the general availability or integrity of the public core of the internet'.[92] At the time of writing, the Paris Call website lists 76 States (and a large number of NGOs, think tanks, private sector companies etc.) as supporters of the Call.[93] While signing the Paris Call cannot be understood as evidence of *opinio iuris* for the existence of an obligation to prevent activity damaging the availability or integrity of the Public Core, it shows that there is increasing understanding of the internet as a common good and the need to protect its critical functions.

Finally, the concept of the public core of the internet has already found its way into legislation. On 17 April 2019, the European Parliament and the Council adopted Regulation (EU) 2019/881, better known as the EU Cybersecurity Act.[94] In Recital 23, the Regulation stipulates that the 'public core of the open internet, namely its main protocols and infrastructure' are a global public good.[95] To protect this public good, the European Union Agency for Cybersecurity (ENISA) shall '[assist] Member States and Union institutions, bodies, offices and agencies in developing and promoting cybersecurity policies related to sustaining the general availability or integrity of the public core of the open internet'.[96]

2) Elements of the Public Core

As the concept of the public core is still under development, its elements are not yet fully defined. The Netherlands Scientific Council report limited it to the logical and physical layers of the internet as a deliberate 'lowest common denominator' approach to secure as much international support as possible for a norm to protect the core from malicious interference.[97] At a minimum, this would include those elements of the logical layer (TCP/IP, DNS, routing protocols etc.), the physical layer (DNS servers,

90 Global Commission on the Stability of Cyberspace, 'Call to Protect the Public Core of the Internet' (2017) https://cyberstability.org/research/call-to-protect/ [19.04.2020].
91 'Paris Call for Trust and Security in Cyberspace' (2018) https://pariscall.international/en/call [19.04.2020].
92 Ibid.
93 https://pariscall.international/en/supporters [04.01.2020].
94 Regulation (EU) 2019/881 of the European Parliament and of the Council 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), OJ L 151, 7.6.2019, 15–69.
95 EU Cybersecurity Act, rec 23.
96 EU Cybersecurity Act, Art. 5.
97 Broeders (n 87) 2.

sea cables) and an organisational layer (internet exchanges, CERTs), which are necessary to ensure the proper functioning of the global internet from a technological standpoint.[98] Similarly, the GCSC Call and Final Report defined the concept of the public core as including 'such critical elements of the infrastructure of the internet as packet routing and forwarding, naming and numbering systems, the cryptographic mechanisms of security and identity, transmission media, software, and data centers'.[99] The Paris Call for Trust and Security in Cyberspace did not specify the elements of the public core of the internet, but it is clear from the accompanying examples given on the official website that it should include the Domain Name System and other critical protocols.[100] More helpful in this regard is the EU Cybersecurity Act, which includes in the public core key protocols (such as DNS, BGP and IPv6), the operation of the domain name system and the operation of the root zone.[101]

In consequence, while the concept is still evolving and despite remaining uncertainties, it seems clear that there is growing consensus that the public core of the internet should at least include the key protocols, the domain name system and the root zone, as described in the EU Cybersecurity Act.

3) Towards an International Collective Obligation to Protect the Public Core?

In the opinion of the present author, the obligation to protect the public core of the internet is a good candidate for a cyber-specific community interest norm. The proper functioning of the public core affects the international community because an attack against the DNS system or key internet protocols would affect every State with an internet connection. By its design and intended function, the obligation to protect the public core is not concerned with the rights of individual States, but rather with the proper functioning of a common good. For these reasons, all States would have an interest in the protection of the public core.

Of course, we are still a long way from the protection of the public core of the internet becoming a legal obligation of *erga omnes (partes)* character. However, as the Paris Call shows, there is international momentum acknowledging and supporting the need to set up a norm protecting it, and in the EU Cybersecurity Act, the first legislative steps have been taken. It is, therefore, conceivable that this momentum will generate further steps to first acknowledge the existence of a soft-law 'cyber norm' to protect the public core. The current deliberations of the UN Group of Governmental Experts and the Open-ended Working Group seem encouraging for such a step. Once the obligation to protect the public core gains recognition within the UN system, it might then follow the path taken by some environmental norms. For instance, the 1992

98 Cf. Ibid. 3.

99 Global Commission on the Stability of Cyberspace, 'Advancing Cyberstability' (2019) https://cyberstability.org/report/, Appendix B, Norm Nr. 1.

100 https://pariscall.international/en/principles [06.01.2020].

101 EU Cybersecurity Act, rec 23.

Rio Declaration[102] contains as non-binding principles the obligation to undertake environmental impact assessments (Principle 17) and the principle of sustainable development (Principle 4). Due to their proliferation in international treaties, soft law and national legislation, the International Court of Justice has defined the obligation to undertake an environmental impact assessment as a 'requirement under general international law'[103] and has applied the principle of sustainable development as one of the factors in interpreting environmental treaties.[104] The obligation to protect the public core of the internet might follow the same route.

5. CONCLUSIONS AND OUTLOOK

This analysis has shown that under current international law, States which fall victim to cyberattacks may count on collective support in two circumstances: where the cyberattack in question was sufficiently grave to constitute an armed attack so that other States may take action in collective self-defence, or where collective reactions are confined to actions which themselves do not amount to violations of international law. It is thus permissible for non-injured States to apply travel bans and asset freezes against individual perpetrators, but not to take offensive action in the networks of the responsible State if that action would violate the principle of non-intervention or that State's sovereignty. However, international law is not static, but rather constantly changing and developing and the norm to protect the public core of the internet might and should evolve into a legally binding community norm, and all States would have a legal interest in its protection.

Apart from that, would the progressive development of international law to allow collective countermeasures in cyberspace against violations of any norm of international law be a good idea? There are certainly sound policy arguments which might support this proposition.[105] First and foremost, the current legal regime limits the options for helping States facing even large-scale cyberattacks. If a State does not possess autonomous offensive cyber capabilities and other States are not allowed to conduct offensive cyber operations as third-party countermeasures, hacking back against the perpetrators of the attacks would either be impossible for the affected State (due to a lack of capabilities) or legally impermissible for third States possessing the necessary capabilities and willing to help. This might lead to a pressure on all States to acquire offensive cyber capabilities and in the meantime restrict the affected State to resort to slower, not in-kind countermeasures.[106] Additionally, it might lead States to deny the applicability of the obligation to respect the territorial sovereignty

102 Rio Declaration on Environment and Development (1992), A/CONF.151/26, vol I.
103 *Pulp Mills on the River Uruguay* (Argentina v Uruguay), Judgment, [2010] ICJ Rep 14, para 204.
104 Ibid. para 177.
105 Similarly Michael N Schmitt, 'Estonia Speaks Out on Key Rules for Cyberspace' (*Just Security*, 10 June 2019) https://www.justsecurity.org/64490/estonia-speaks-out-on-key-rules-for-cyberspace/ [19.04.2020].
106 Ibid.

of a State in cyberspace to avoid simple hack-back cyber operations being qualified as violations of sovereignty and thus internationally wrongful acts, thereby avoiding the need for their justification as collective countermeasures.[107] Finally, allowing collective countermeasures against violations of sovereignty or non-intervention in cyberspace would better take into account the specificity of cyber operations, in particular their clandestine nature. In such cases, a hack-back against the source of the cyberattack is often the most direct and effective way to cause the attacking State to stop the cyber operation by disabling the source of the threat, which is the idea of countermeasures in the first place.

In any case, it has to be concluded that Estonia has started a much needed and important discussion among States and scholars, for which it has to be congratulated. States should now – like France – take up this challenge and declare their position towards collective countermeasures. The UN GGE and OEWG would be good venues for such declarations.

ACKNOWLEDGEMENTS

I would like to thank Dr Dennis Broeders and the team of the Hague Programme for Cyber Norms at Leiden University for the fruitful discussions about the ideas contained in this paper during my stay as Visiting Fellow of the Cyber Norms Programme in September 2019.

REFERENCES

Alland D, 'Countermeasures of General Interest' (2002) 13 European Journal of International Law 1221

Anzilotti D, *Cours de Droit International* (Recueil Sirey 1929)

Arcari M, 'International Reactions to the Crimea Annexation under the Law of State Responsibility: 'Collective Countermeasures' and Beyond?' in Władysław Czapliński and others (eds), *The Case of Crimea's Annexation under International Law* (Scholar 2017)

Bílková V, 'The Use of Force by the Russian Federation in Crimea' (2015) 75 Zeitschrift für ausländisches öffentliches Recht und Völkerrecht 27

Broeders D, *The Public Core of the Internet* (Amsterdam University Press 2015)
——, 'Aligning the International Protection of 'the Public Core of the Internet' with State Sovereignty and National Security' (2017) 2 Journal of Cyber Policy 366

107 Cf. Paul C. Ney, *DOD General Counsel Remarks at U.S. Cyber Command Legal Conference*, Speech By DOD General Counsel Paul C. Ney on 2 March 2020, https://www.defense.gov/Newsroom/Speeches/Speech/Article/2099378/dod-general-counsel-remarks-at-us-cyber-command-legal-conference/ [19.04.2020] (arguing that 'international law [does not generally prohibit] non-consensual cyber operations in another State's territory').

Calamita NJ, 'Sanctions, Countermeasures, and the Iranian Nuclear Issue' (2009) 42 Vanderbilt Journal of Transnational Law 1393

Darcy S, 'Retaliation and Reprisal' in Marc Weller (ed), *The Oxford Handbook of the Use of Force in International Law* (Oxford University Press 2015)

Dawidowicz M, 'Public Law Enforcement without Public Law Safeguards? An Analysis of State Practice on Third-Party Countermeasures and Their Relationship to the UN Security Council' (2010) 77 British Yearbook of International Law 333
——, Third-Party Countermeasures in International Law (Cambridge University Press 2017)

Eichensehr KE, 'The Law & Politics of Cyberattack Attribution' (2019) 19–36 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3453804>

Feichtner I, 'Community Interest', *Max Planck Encyclopaedia of Public International Law* (Oxford University Press 2007)

Focarelli C, 'International Law and Third-Party Countermeasures in the Age of Global Instant Communication' (2016) 29 Questions of International Law 17 <http://www.qil-qdi.org/international-law-third-party-countermeasures-age-global-instant-communication/>

Giegerich T, 'Retorsion', *Max Planck Encyclopaedia of Public International Law* (Oxford University Press 2011)

Global Commission on the Stability of Cyberspace, 'Call to Protect the Public Core of the Internet' <https://cyberstability.org/research/call-to-protect/>
——, 'Advancing Cyberstability' (2019) <https://cyberstability.org/report/>

Katselli Proukaki E, The Problem of Enforcement in International Law (Routledge 2010)

Koskenniemi M, 'Solidarity Measures: State Responsibility as a New International Order?' (2002) 72 *British Yearbook of International Law* 337

Martin Dawidowicz, 'Third-Party Countermeasures: A Progressive Development of International Law? - QIL QDI' (2016) 29 Questions of International Law 3 <http://www.qil-qdi.org/third-party-countermeasures-progressive-development-international-law/>

Marxsen C, 'The Crimea Crisis: An International Law Perspective' (2014) 74 Zeitschrift für ausländisches öffentliches Recht und Völkerrecht 367

Pellet A, 'The Definition of Responsibility in International Law' in James Crawford and others (eds), *The Law of International Responsibility* (Oxford University Press 2010)

Pomerlau M, 'Two Years in, How Has a New Strategy Changed Cyber Operations?' (*Fifth Domain*, 11 November 2019) <https://www.fifthdomain.com/dod/2019/11/11/two-years-in-how-has-a-new-strategy-changed-cyber-operations/>

Ruffert M, 'Reprisals', Max Planck Encyclopaedia of Public International Law (2015)

Sander B, 'Democracy under the Influence: Paradigms of State Responsibility for Cyber Influence Operations on Elections' (2019) 18 Chinese Journal of International Law 1

Schmitt MN, 'Estonia Speaks Out on Key Rules for Cyberspace' (*Just Security*, 10 June 2019) <https://www.justsecurity.org/64490/estonia-speaks-out-on-key-rules-for-cyberspace/>

Schmitt MN and Vihul L, 'International Cyber Law Politicized: The UN GGE's Failure to Advance Cyber Norms' (*Just Security*, 30 June 2017) <https://www.justsecurity.org/42768/international-cyber-law-politicized-gges-failure-advance-cyber-norms/>
—— (eds), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press 2017)

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Up in the Air: Ensuring Government Data Sovereignty in the Cloud

Neal Kushwaha
Founder and Advisor
IMPENDO Inc.
Ottawa, Canada
neal@impendo.com

Bruce W. Watson
Chief Scientist and Advisor
IP Blox and IMPENDO Inc.
Eindhoven, Netherlands
Ottawa, Canada
bruce@ip-blox.com
bruce@impendo.com

Przemysław Roguski
Lecturer
Chair for Public International Law
Jagiellonian University
Kraków, Poland
przemyslaw.roguski@uj.edu.pl

Abstract: Governments around the world commonly use Cloud Service Providers (CSPs) that are headquartered in other nations. How do they ensure data sovereignty when these CSPs, storing a nation's data within that nation's borders, are subject to long-arm statutes on data stored abroad? And what if, in turn, the governmental data is stored abroad, would access to that data constitute a violation of the nation's sovereignty?

This paper examines how selected governments have protected their CSP-hosted data from foreign law enforcement access and suggests methods that other governments might employ to ensure data sovereignty. It addresses these issues in three steps. First, we describe the problem of long-arm jurisdiction with respect to the US Clarifying Lawful Overseas Use of Data (CLOUD) Act and the proposed EU *e-evidence* regulation (EU COM/2018/225 final). Given the extraterritorial reach of these regulations, foreign CSPs looking to maintain good standing with their respective governments and laws may consider storing active copies of their customers' data

and metadata in their home country. For CSPs offering services to the EU, having to comply with an emergency Production Order within 6 hours may not be possible without duplication and active parsing of customer data in a CSP centralised location, foreign to the data owner.

Secondly, we evaluate the recently signed US and UK Executive Agreement under the US CLOUD Act to see if and how the UK protects its own Government Cloud from US law enforcement. We also evaluate France's position, the German model which prohibits storing of government data with US CSPs, and the Polish model recently signed with the US CSP Google, to better understand their positions and approach to managing data sovereignty.

Finally, we offer an assessment on how to balance sovereignty over government data stored in the cloud with the needs of law enforcement for States exercising jurisdiction over CSPs.

Keywords: *cloud, data sovereignty, international law*

1. INTRODUCTION

The classical notion of sovereignty, dating back to the 16th century, signifies the highest authority of a State and the right to exercise its own judgment within a territory.[1] Internally, it denotes the State's exclusive competence to enact and enforce laws binding on its territory and to decide freely in all internal matters not regulated by international law, including the right to control access to its territory.[2] Consequently, violations of sovereignty under international law include violation of territorial integrity and impacting inherently governmental functions. Breaches of sovereignty via cyber means may not be as clear. In July 2015, the UN Group of Governmental Experts (GGE) reached consensus confirming that sovereignty applies to the conduct by States of Information and Communications Technology (ICT) related activities and to their jurisdiction over ICT infrastructure within their territory.[3]

Another aspect of sovereignty is jurisdiction, i.e. the State's right under international law to regulate conduct in matters not exclusively of domestic concern.[4] The GGE

1 PCIJ, *Customs Régime between Germany and Austria (Protocol of March 19th, 1931)*, Advisory Opinion, 1931 PCIJ Series A/B No 41, Sep. opinion Judge Anzilotti, para 13.
2 Samantha Besson, 'Sovereignty' in Rüdiger Wolfrum (ed), *Max Planck Encyclopaedia of Public International Law* (Oxford University Press 2011) para 118ff.
3 UN Doc. A/70/174 paras 26, 27, and 28(b), 'How International Law applies to the use of ICTs'.
4 F.A. Mann, 'The Doctrine of Jurisdiction in International Law', Recueil des Cours de l'Académie de Droit Internationale, 111 (1964), 2.

confirmed in its 2015 Report that 'States have jurisdiction over the ICT infrastructure located within their territory'.[5] However, the GGE did not form a consensus view as to jurisdiction over data stored within that ICT infrastructure. This unresolved question causes significant practical problems.

Consider the following example: a governmental department in State A extends a contract to a Cloud Service Provider (CSP) headquartered in State B. The State A department consumes various cloud services, storing transactional data (emails, calendars, etc.) and master data (names, addresses, social insurance numbers, income, etc.) in the CSP's data centres located in State A and replicated in other data centres under the CSP's control, specifically those in State B. Consider further that State B may authorise its law enforcement agencies to access all data stored by CSPs registered in State B and oblige those CSPs to preserve and hand over the data on production of a warrant. Alternatively, the data in the State B data centre is breached by foreign non-state actors and copied to data centres located in State C. The non-state actors then publicly share the content from the State C data centres for anyone to consume, exposing conversations, contacts and State A citizen data, risking identity theft, cybercrimes and more.

In both examples, State A loses control over its data, at the same time enabling another sovereign (in variant 1) to exercise control over that data by virtue of it being stored in that State's territory. This article explores the concept of data sovereignty and asks whether unconsented access to government data stored abroad would constitute a violation of a State's sovereignty; and what States can do to ensure continued control over their data.

This is not only a theoretical problem. At the time of writing, the Government of Canada[6] (GC) has been consuming cloud services for over 7 years.[7] Services such as those provided by Google, Microsoft, Amazon, ServiceNow, and Salesforce are currently being used for production workloads. Various Canadian departments, agencies, crown corporations, tribunals, etc. (herein collectively called departments) use cloud capacities in a variety of ways. Some have email services fully hosted in the cloud, while others use unique services in conjunction with existing internal services.[8]

5 UN Doc. A/70/174 para 28(a).

6 By virtue of the origin and experience of two of the authors, Canada is used throughout the text as an example of a government that may benefit from reviewing its approach to the cloud through the lens of international law and its applicability to governmental data stored domestically or abroad.

7 See Jean-Martin Thibeault, 'This just in! Canadian Broadcasting Corporation moves 12,000 accounts to Google Apps in 90 days' (Google Cloud Official Blog, 14 May 2013), https://cloud.googleblog.com/2013/05/this-just-in-canadian-broadcasting.html [14.04.2020]. Canadian Broadcasting Corporation (CBC) is a Canadian Crown Corporation and is accountable to the Canadian Parliament.

8 Naming these departments and describing the respective services they consume is not the purpose of this paper.

Large US CSPs such as Microsoft and Amazon Web Services commonly operate and offer their services in Canada under a Canadian corporation. Both Microsoft Canada and Amazon Web Services Canada now offer their Canadian customers the option to host their data in Canadian data centres. Although a Canadian citizen, corporation or the GC may consider this adequate, with the enacting into law of the US Clarifying Lawful Overseas Use of Data Act (CLOUD Act),[9] these Canadian entities may not only be exposing their data to various law enforcement agencies within the US, but also to other States holding executive agreements with the US.

The problems described above are universal, though, and other countries are also grappling with them. How do nations consuming (or intending to consume) CSP services ensure data sovereignty while the nations where the CSPs are headquartered enact new laws?

This paper begins by introducing the concept of 'data sovereignty' and describing some of the implications of long-arm jurisdictions for Internet Service Providers (ISPs) and CSPs and how they will likely manage the tight timelines for evidence requests. It then performs a high-level evaluation of the US and UK Executive Agreement,[10] France's position, the Domestic Cloud Provider (DCP) agreement of Chmura Krajowa (under the Polish Development Fund and PKO Bank Polski) to manage and resell Google's cloud offering,[11] and Germany's GAIA-X project.[12]

We close the paper with suggestions for nations to move forward with consuming cloud services while ensuring data sovereignty.

2. THE CONCEPT OF DATA SOVEREIGNTY

The concept of data sovereignty has been the topic of scholarly debate for some time,[13] but has recently gained traction within States as well.[14] It is not yet a fully

9 Consolidated Appropriations Act of 2018, Pub. L. No. 115-141, 132 Stat. 348, div. 5 (2018).

10 Agreement on Access to Electronic Data for the Purpose of Countering Serious Crime [CS USA No.6/2019], https://www.gov.uk/government/publications/ukusa-agreement-on-access-to-electronic-data-for-the-purpose-of-countering-serious-crime-cs-usa-no62019 [14.04.2020], hereinafter '*US-UK Agreement*'.

11 Operator Chmury Krajowej, *Press release of 27 September 2019*, https://chmurakrajowa.pl/partnership.html [14.04.2020].

12 Bundesministerium für Wirtschaft und Energie, *Das Projekt GAIA-X, Eine vernetzte Dateninfrastruktur als Wiege eines vitalen, europäischen Ökosystems*, https://www.bmwi.de/Redaktion/DE/Publikationen/Digitale-Welt/das-projekt-gaia-x.html, also available in English: https://www.bmwi.de/Redaktion/EN/Publikationen/Digitale-Welt/project-gaia-x.html [14.04.2020].

13 Jing de Jong-Chen, 'Data Sovereignty, Cybersecurity and Challenges for Globalization', Georgetown Journal of International Affairs, (Fall 2015), 112-122; Patrik Hummel, Matthias Braun et al., 'Sovereignty and Data Sharing', ITU Journal: ICT Discoveries, Special Issue No. 2; Andrew Keane Woods, 'Litigating Data Sovereignty' (2018) 128 Yale Law Journal 328.

14 Bundesministerium für Wirtschaft und Energie, *Das Projekt GAIA-X, Eine vernetzte Dateninfrastruktur*, 6.

settled term and States use it in conjunction with other notions such as technological sovereignty[15] or digital sovereignty.[16] For the purposes of this paper, it is best to look at these concepts as akin to a Russian doll, where the broadest concept, technological sovereignty, encompasses the narrower digital sovereignty, which in turn encompasses data sovereignty. Technological sovereignty has been described by European Commission President Ursula von der Leyen as Europe's capability 'to make its own choices, based on its own values, respecting its own rules' in the field of tech.[17] It includes, amongst others, the integrity and resilience of data infrastructure, networks and communications[18] and the development of autonomous capacities in the field of artificial intelligence.[19] The slightly narrower term of 'digital sovereignty' (or '*souveraineté numérique*') has been gaining popularity mainly in France, where it was first introduced by Pierre Bellanger, president of Skyrock,[20] and since then taken up by State organs such as the French Senate,[21] but has been also used by other States such as Germany. In the German view, 'digital sovereignty' (or '*digitale Souveränität*') denotes the 'capability to take autonomous actions and decisions in the digital environment'.[22] The French view is similar and refers to the application of the principle of sovereignty to cyberspace and includes, amongst other aspects relating to the ability to detect and react to threats in cyberspace,[23] control over data in cyberspace.[24]

15 European Commission, *Shaping Europe's Digital Future*, February 2020, 3, doi:10.2759/091014 [14.04.2020]; for a discussion of the term 'technological sovereignty' see also Tim Maurer et al., 'Technological Sovereignty: Missing the Point?', in: M.Maybaum, A.-M.Osula, L.Lindström (Eds.), *2015 7th International Conference on Cyber Conflict:Architectures in Cyberspace*, (NATO CCDCoE Publications 2015), 53ff.

16 French Ministry for Europe and Foreign Affairs, *Intervention de Jean-Yves Le Drian, ministre de l'Europe et des Affaires étrangères, au colloque 'Au-delà de 1989 Espoirs et désillusions après les révolutions*', Speech by Minister Jean-Yves Le Drian in Prague on 6 December 2019, https://www.diplomatie.gouv.fr/fr/dossiers-pays/republique-tcheque/evenements/article/intervention-de-jean-yves-le-drian-ministre-de-l-europe-et-des-affaires [14.04.2020] (referring to the need to construct 'European digital sovereignty' (*souveraineté numérique européenne*)).

17 Ursula von der Leyen, 'Tech Sovereignty Key for EU's Future Goals', *The Irish Examiner* (Cork, 18 February 2020), https://www.irishexaminer.com/breakingnews/views/analysis/ursula-von-der-leyen-tech-sovereignty-key-for-eus-future-goals-982505.html [14.04.2020].

18 European Commission, *Shaping Europe's Digital Future*, February 2020, 3, DOI:10.2759/091014 [14.04.2020].

19 European Commission, *Press remarks by President von der Leyen on the Commission's new strategy: Shaping Europe's Digital Future*, 19 February 2020, https://ec.europa.eu/commission/presscorner/detail/en/speech_20_294 [14.04.2020].

20 Pierre Bellanger, 'De la souveraineté en général et de la souveraineté numérique en particulier', *Les Echos* (30 August 2011), http://archives.lesechos.fr/archives/cercle/2011/08/30/cercle_37239.htm [14.04.2020]; Pierre Bellanger, La Souveraineté Numérique (Stock 2014).

21 Commission d'enquête sur la souveraineté numérique, *Rapport de Gérard LONGUET sur la souveraineté numérique, fait au nom de la commission d'enquête*, Rapport n° 7 (2019-2020), (Report, 1 October 2019), http://www.senat.fr/rap/r19-007-1/r19-007-1.html [14.04.2020], hereinafter Rapport n° 7.

22 Bundesministerium für Wirtschaft und Energie, 'Digitale Souveränität im Kontext plattformbasierter Ökosysteme', 6, https://www.de.digital/DIGITAL/Redaktion/DE/Digital-Gipfel/Download/2019/digitale-souveraenitaet.pdf [14.04.2020].

23 French Ministry of the Armies, 'International Law Applied to Operations in Cyberspace', https://www.defense.gouv.fr/content/download/567648/9770527/file/international+law+applied+to+operations+in+cyberspace.pdf [14.04.2020].

24 Rapport n° 7, 17.

Therefore, the issue of control is the heart of the concept of data sovereignty.[25] In its classical form, sovereignty denotes supreme authority over territory to the exclusion of other sovereigns.[26] Within the specified territory, the sovereign has the exclusive power to exercise its own judgment.[27] It can control access to its territory and enact and enforce laws with respect to persons and objects within this territory. This traditional understanding is challenged by data in cyberspace. As it is only bits and bytes, which can be moved instantaneously across borders, copied, stored in multiple locations and split into parts, while remaining accessible from within the territory, data becomes 'un-territorial'.[28] Given that territoriality loses its importance with respect to data in cyberspace, the main aspect of sovereignty over data becomes the exclusive authority or control: 'sovereign data subjects are those who are in a position to articulate and enforce claims to power about their data'.[29] Consequently, the concept of data sovereignty denotes exclusive control over stored and processed data and the ability to decide who is granted access to that data.[30]

With respect to the topic of this paper, data sovereignty is particularly relevant in the context of cloud computing. Given that governments are increasingly moving their services and data, including both governmental and citizens' data, to the cloud, *who* exercises control over it is a question of sovereignty. As many of the largest CSPs are located in the United States and CSPs are currently free to store data in offshore data centres, governmental data could be stored on servers within the territories of several States and be subject to the jurisdiction, and thus sovereign control, of each of those States. Thus, competing jurisdictions and control over network infrastructure (data centres) and CSPs directly challenge the exclusive control a government may expect over its data, and States' sovereignty over their data in general.[31]

In the next two sections, we will discuss how States gain control over data through long-arm jurisdiction over CSPs residing or operating within the territory of those States and how other States react to meet this challenge and protect their data sovereignty.

25 See, e.g. Bundesministerium für Wirtschaft und Energie, *Das Projekt Gaia-X: Eine vernetzte Dateninfrastruktur als Wiege eines vitalen, europäischen Ökosystems*, 15, https://www.bmwi.de/Redaktion/DE/Publikationen/Digitale-Welt/das-projekt-gaia-x.pdf [14.04.2020], defining 'data sovereignty' as 'guarantee of control over the use of data' ('*Garantie der Datennutzungskontrolle*').

26 Samantha Besson, 'Sovereignty' in Rüdiger Wolfrum (ed.), *Max Planck Encyclopaedia of Public International Law* (Oxford University Press 2011) para 118ff.

27 PCIJ, *Customs Régime between Germany and Austria (Protocol of March 19th, 1931)*, Advisory Opinion, 1931 PCIJ Series A/B No 41, sep. opinion Judge Anzilotti at para 13.

28 Jennifer Daskal, 'Borders and Bits' (2018) 17 Vanderbilt Law Review 179, 181; see also Jennifer Daskal, 'The Un-Territoriality of Data' (2015) 125 Yale Law Journal 326.

29 Patrik Hummel, Matthias Braun et al., 'Sovereignty and Data Sharing', ITU Journal: ICT Discoveries, Special Issue No. 2, 2.

30 Bundesministerium für Wirtschaft und Energie, *Digitale Souveränität im Kontext plattformbasierter Ökosysteme*, .6, https://www.de.digital/DIGITAL/Redaktion/DE/Digital-Gipfel/Download/2019/p2-digitale-souveraenitaet-plattformbasierter-oekosysteme.pdf [14.04.2020].

31 For a broader discussion of this argument, see Andrew Keane Woods, 'Litigating Data Sovereignty' (2018) 128 Yale Law Journal 328, 360 ff.

3. LONG-ARM JURISDICTIONS

Both the US CLOUD Act and the draft EU Regulation on European Production and Preservation Orders for electronic evidence[32] allow for the preservation and production of data stored by a service provider in another jurisdiction as evidence in criminal investigations.

The US CLOUD Act (title 18 U.S.C. §2713) allows for extraterritorial reach of all US CSP data. Under 'Required preservation and disclosure of communications and records', it specifically states:

> 'A provider of electronic communication service or remote computing service shall comply with the obligations of this chapter to preserve, backup, or disclose the contents of a wire or electronic communication and any record or other information pertaining to a customer or subscriber within such provider's possession, custody, or control, regardless of whether such communication, record, or other information is located within or outside of the United States'.

Specific in its timelines, the EU COM/2018/225 final proposal requires Member States to respond to requests within 10 days for standard requests and 6 hours in an emergency.[33] The previous response times were on average 10 months for Mutual Legal Assistance and 120 days for European Investigation Orders.[34] The Production Orders and Preservation Orders relate to four data types listed in the proposal and apply to any service provider, regardless of where the parent company is located or where the data is held.[35]

1. **Subscriber data:** personal information used to identify an individual, commonly considered Protected B[36] information at the GC level, including

32 Proposal for a Regulation of the European Parliament and the Council on European Production and Preservation Orders for electronic evidence in criminal matters, EU COM/2018/225 final - 2018/0108 (COD).

33 Ibid. Article 9 'Execution of an EPOC', all paragraphs describe the deadlines and how to respond.

34 'What will the new rules change?' in the European Commission's *Frequently Asked Questions: New EU rules to obtain electronic evidence*, 17 April 2018, https://ec.europa.eu/commission/presscorner/detail/en/MEMO_18_3345 [14.04.2020].

35 EU COM/2018/225 final. Article 2 'Definitions', paragraphs 6-10 define and describe electronic evidence as four data types.

36 In Canada, the compromise of 'Protected' information or assets could cause various levels of injury to a non-national interest. The Protected levels are described as A, B, and C. The compromise of 'Classified' information or assets could cause various levels of injury to national interests. The Classified levels are Confidential, Secret and Top Secret. See Treasury Board of Canada Secretariat, *Directive on Security Management - Appendix J: Standard on Security Categorization*, 01 July 2019, https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32614 [14.04.2020]. It should be noted that certain Canadian Crown Corporations (e.g. Bank of Canada) create their own security categorisations which do not align with those described in Appendix J. See Treasury Board of Canada Secretariat, *List of Crown corporations*, 02 February 2019, https://www.canada.ca/en/treasury-board-secretariat/services/guidance-crown-corporations/list-crown-corporations.html [14.04.2020].

name, address, billing information, date of birth, email address, telephone number, etc.

2. **Access data:** a component of metadata, including the logon and log-off dates and times, IP addresses assigned by service providers, etc. It is common to all internet users and generally available from ISPs, and includes IP addresses assigned to the user and the connection times.
3. **Transactional data:** a component of metadata, including geolocation of the source and destination of the data, size of data, route, communication protocol, etc., available from the ISP and CSP. The ISP is able to describe the route and the internet services the user has consumed (examples: access to websites, use of encryption, download of data streams, etc.) while the CSP is able to describe the opening and closing times of a document stored at the CSP, the length of time spent composing an email over the CSP infrastructure, the recipients of the email, and more, but not the content of the document or email.[37]
4. **Content data:** the digital data consumed by the user in voice, video, audio, text, images, etc.

From the perspective of managing criminal proceedings, the US CLOUD Act and EU proposal make it easier to quickly request and gather electronic evidence, however, this places a burden on CSPs. Being prepared to respond to a potential 6-hour or 10-day response for a Production Order likely means the CSP will invest in resources (including staff, equipment and software) to help manage these quick turn-around requests. This investment is unlikely to be in each country in which it operates and will probably be in a central location, at least for the smaller CSPs.

CSPs and ISPs are often not the same company and that likely means the request must go to at least two different parties. Furthermore, users often store information at more than one CSP; for example, data may be stored on Google Drive and in Amazon containers while using email services from Microsoft Office 365. It is thus possible that a Production Order may reach multiple companies.

By way of an example, Canadian ISPs already have the responsibility to provide basic metadata data to law enforcement agencies when lawfully requested.[38] Besides their customers' online activity, ISPs maintain records of user logon and log-off dates and times along with associated IP addresses. This log data is very simple to parse, join with customer account datasets and use for reporting.

On the other hand, due to the volume and velocity of log data, CSPs looking to remain in good standing with their respective governments and laws may consider keeping active copies of their customers' metadata (activity logs) and possibly their customers'

37 Content data (data type #4) may be encrypted and inaccessible to the CSP.
38 See Canada's Personal Information Protection and Electronic Documents Act, SC 2000, c 5, s 7(3)(c).

data in their headquarters country for easy and centralised parsing. For those offering services to the EU, having to comply with an emergency Production Order in 6 hours may not be possible without duplication and active parsing of customer data and/or metadata.

The duplication of data into a CSP's data centre located in another country may result in a breach of sovereignty, while the constant parsing of customer data and metadata may be considered a breach of privacy. CSPs will need to remain aware of laws in the various States they operate in, and international law, all at the same time. This complicated legal knowledge coupled with the required technical knowledge may be economically unreasonable for smaller CSPs and start-ups.

4. INTERNATIONAL POSITIONS

A. US-UK Executive Agreement

The first international cooperation agreement (Agreement) under the US CLOUD Act was signed on 3 October 2019 between the United States and the United Kingdom.[39] Its aim is to create a mechanism for cooperation in allowing law enforcement agencies access to data stored outside their territory and held by service providers registered in a State party to the agreement. In this regard, it will be most beneficial to the UK, allowing it to make direct requests to American CSPs and therefore overcoming the blocking provisions of the US Stored Communications Act, which had previously prevented American CSPs from handing over data stored in the US to foreign law enforcement agencies.[40] In short, the Agreement allows the parties to directly request from CSPs stored content data, traffic data or metadata, subscriber information and intercept wire electronic communications related to a serious crime investigation (Article 1(3)).

Production Orders may be addressed directly to the Covered Providers (i.e. any private entity which provides to the public the ability to communicate or to process or store computer data, by means of a Computer System or a telecommunications system; or processes or stores Covered Data, Article 1(7)) and the Providers have to produce the information directly to designated authorities of each Party (Article 10).

The Agreement includes limitations on the use and transfer of data (Article 8) and privacy and data protection safeguards (Article 9). Both parties certify that their legal

39 US-UK Agreement [CS USA No.6/2019].

40 Theodore Christakis, *21 Thoughts and Questions about the UK-US CLOUD Act Agreement: (and an Explanation of How it Works – with Charts)*, https://europeanlawblog.eu/2019/10/17/21-thoughts-and-questions-about-the-uk-us-cloud-act-agreement-and-an-explanation-of-how-it-works-with-charts/ [14.04.2020]; see also Jennifer Daskal, Peter Swire, 'The UK-US CLOUD Act Agreement is Finally Here, Containing New Safeguards', (*Just Security*, 08 October 2019), https://www.justsecurity.org/66507/the-uk-us-cloud-act-agreement-is-finally-here-containing-new-safeguards/ [14.04.2020].

systems have adequate protections for privacy and civil liberties (which is required by the CLOUD Act for the US to enter into such an agreement) and the Agreement establishes procedural requirements and oversight mechanisms to comply with privacy requirements.[41] Under Article 5(10) of the Agreement, the party issuing a Production Order is obliged to notify the authorities of a third country, where an 'Order subject to this Agreement is issued for data in respect of an individual who is reasonably believed to be located outside the territory of the Issuing Party and is not a national of the Issuing Party'. This is with the exception of cases where notification would endanger national security or the notification would imperil human rights.

Three inherent limitations of the US-UK Agreement need to be stressed. *Firstly*, pursuant to Article 6(3) of the Agreement, it 'does not in any way restrict or eliminate any legal obligation Covered Providers have to produce data in response to Legal Process issued pursuant to the law of the Issuing Party'. Consequently, Article 6(3) does not exclude the parallel application of national provisions such as the CLOUD Act to data of persons covered by the Agreement. Thus, US law enforcement can request the production of data of UK (or EU) persons held by CSPs falling under the CLOUD Act, without necessarily having to fulfil other obligations under the Agreement, such as the notification of third parties.[42] Therefore, it is not clear under which incentives US law enforcement would choose to use the Agreement, rather than the CLOUD Act.[43]

Secondly, the Agreement authorises both parties to request from CSPs data of persons residing in a third country (such as Canada or the European Union Member States, for example), provided it is stored in the territory of the parties.[44] In effect, this might create conflicts with third States, which would presumably not take kindly to such practices or might also resort to applying such measures themselves.

Thirdly, the Agreement does not address the question of governmental data or data of government employees which might be connected with the exercise of their official duties and thus affect not only the interests of those individuals, but also the sovereign interests of the State. Thus, if an American CSP holds data under a contract with a British governmental agency, the CLOUD Act remains potentially applicable and the CSP might still be required to hand over such data to American law enforcement if duly ordered to do so.

41 Nathan Swire, 'Applying the CLOUD Act to the U.S.-UK Bilateral Data Access Agreement', (*Lawfare*, 28 October 2019), https://www.lawfareblog.com/applying-cloud-act-us-uk-bilateral-data-access-agreement [14.04.2020].

42 Theodore Christakis, '21 Thoughts and Questions about the UK-US CLOUD Act' (*European Law Blog*, 17 October 2019), https://europeanlawblog.eu/2019/10/17/21-thoughts-and-questions-about-the-uk-us-cloud-act-agreement-and-an-explanation-of-how-it-works-with-charts/ [14.04.2020].

43 Ibid.

44 Ibid.

In sum, the US-UK Executive Agreement on access to electronic data to combat serious crime is a welcome step to address issues of extraterritorial data Production Orders with a dedicated international legal instrument. Nevertheless, the subjection of the Agreement to domestic provisions such as the CLOUD Act raises certain questions as to its status under international law.[45] For the purposes of this analysis, it is important to note that the main issue with governmental cloud systems depending on the services of foreign CSPs is not addressed. The Agreement governs only Production Orders for data of individuals that is needed to combat serious crimes, not industrial or governmental data. Neither does it give the UK government a way to address potential Production Orders for governmental data by American law enforcement, be it under the CLOUD Act or other provisions of domestic law, within an agreed international framework. This issue therefore remains unresolved by the Agreement.

B. France

France was one of the pioneers of the notion of digital sovereignty (*souveraineté numérique*) in Europe.[46] In 2019, the French Senate convened an Inquiry Committee (*Commission d'Enquête*) on the topic of digital sovereignty with the aim of studying the issue and formulating policy recommendations. Its final report, presented by Rapporteur Gérard Longuet, critically examined, among others, the question of cloud storage and extraterritorial jurisdiction.[47] It held that in the modern world, data has become an economic strategic issue (*enjeu économique stratégique*) of immense importance to the activities of the major actors of the digital economy.[48] The report discussed the question of data localisation as one of the modes of protecting data, but found it an imperfect solution.[49] It found that data localisation rules might be important with respect to securing digital sovereignty in three instances: in cases of strategic or particularly sensible data such as data pertaining to public finances (*traitements publics souverains*), private financial data or commercial secrets, to guarantee access to essential services and to support the industrial ecosystem of cloud providers.[50]

The report noted, however, that data localisation clauses do not ameliorate the risks posed both by extraterritorial legislation such as the CLOUD Act and the dependence of certain technology companies on their States, as with certain Chinese companies.[51] It criticised the CLOUD Act as being too broad with respect to the affected entities,

45 Ibid.
46 For an overview of French scholarly literature on this matter Pierre Bellanger, *La souveraineté numérique*, (Paris, Stock 2014); Marin Brenac, Pierre-Luc Déziel, *La souveraineté numérique sur les données personnelles : étude du règlement européen no 2016/679 sur la protection des données personnelles à l'aune du concept émergent de souveraineté numérique*, (Québec, Université Laval 2017); Pauline Türk, Christian Vallar, *La souveraineté numérique: le concept, les enjeux*, (Paris, Editions Mare & Martin 2018).
47 Rapport n° 7.
48 Ibid, 54.
49 Ibid, 68.
50 Ibid.
51 Ibid, 69.
52 Ibid, 71.

the infractions covered and the type and amount of data collected[52] and found the CLOUD Act to pose a risk of access by American law enforcement to strategic data of legal persons and to be incompatible with the GDPR with regard to the protection of personal data.[53]

The report discusses three options to mitigate those risks: *firstly*, the legal separation of subsidiary companies for each region and geographical location of services;[54] *secondly*, mobilising companies on a case-by-case basis to contest excessive law enforcement demands in court; and, *thirdly*, the extensive use of robust data encryption technologies.[55] Similar to the report of 26 June 2019, prepared for the French Prime Minister by Raphaël Gauvain,[56] the Longuet Report advises the strengthening of the 1968 law on blocking measures,[57] extending the protections of the GDPR to non-personal data of legal persons and sanctioning their 'improper transmission' (*transmission indue*) and encouraging the fast conclusion of a cooperation agreement between the European Union, its Member States and the US.[58]

While none of these measures has been implemented at the time of writing, what becomes clear from the Gauvain and Longuet reports is that France is deeply concerned about American (and Chinese) extraterritorial reach, brought about by their dominance in the software and hardware sectors, respectively. The French view is that it has to take robust action, both legislative and in terms of industrial policy, to protect French data and French strategic interests against the reach of foreign States, even like-minded States such as the US.

C. Germany

Similar considerations underpin the German position with respect to American cloud services. Ever since the Snowden revelations, Germany has been deeply worried about the access of the US National Security Agency and US law enforcement to German data. The German government has repeatedly stressed that while it recognises the importance of facing up to novel challenges to law enforcement posed by the proliferation of transnational cloud services, any solution needs to respect fundamental

53 Ibid, 72.

54 The report refers as an example to the French company OVH, which has set up a dedicated company for its activities in the United States, presumably to separate the parent company's data from American law enforcement requests.

55 Ibid, 74.

56 Raphael Gauvain, *Rétablir la souveraineté de la France et de l'Europe et protéger nos entreprises des lois et mesures à portée extraterritoriale, Rapport à la demande de Monsieur Édouard Philippe, Premier Ministre*, (Report, 26 June 2019), https://www.vie-publique.fr/sites/default/files/rapport/pdf/194000532.pdf [14.04.2020].

57 Referring here to a 1968 Statute prohibiting, subject to international treaties, the passing on to foreign governments of documents or information of an economic, commercial, industrial, financial or technical nature, the communication of which is likely to undermine France's sovereignty, security, essential economic interests or public order. 'Loi n° 68-678 du 26 juillet 1968 relative à la communication de documents et renseignements d'ordre économique, commercial, industriel, financier ou technique à des personnes physiques ou morales étrangères', https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000501326&categorieLien=cid [14.04.2020].

58 Rapport n° 7, 75.

human rights and facilitate cooperation between States. To this end, it advocates rapid negotiations between the European Commission and the US government to conclude a cooperation agreement on data sharing, as envisaged by the CLOUD Act.[59]

Germany also seeks to secure its digital sovereignty (*digitale Souveränität*), limiting US law enforcement access to German data. This is done via two routes: *firstly*, by limiting the type of data that can be stored on US cloud services; and *secondly*, by developing an autonomous cloud storage solution. It is interesting to note that the first route is driven not only (or even predominately) by the government, but by regional data protection agencies. For instance, the Hessian Commissioner for Data Protection and Freedom of Information (*Der Hessische Beauftragte für Datenschutz und Informationsfreiheit*) has forbidden schools in the Land of Hesse to use Microsoft Office 365 services or to store students' personal data in the cloud if the providers are subject to US law. He said that 'public institutions in Germany have a special responsibility regarding the permissibility and traceability of the processing of personal data [and that] the digital sovereignty of State data processing must also be guaranteed'.[60] The main reason for this statement is that Microsoft, like other CSPs, does not reveal what kind of data is being transmitted to the US and whether US law enforcement would be able to access this data.

To address these concerns, on 29 October 2019, the German government launched the GAIA-X project.[61] The stated motivation for this project is to preserve European 'data sovereignty' (*Datensouveränität*) against increasing dependence on foreign digital technologies.[62] The report defines digital sovereignty as the 'possibility of independent self-determination of State and organisations' with regard to the 'use and design of digital systems themselves, the data generated and stored therein and the processes represented by them'[63] and data sovereignty as 'guarantee of control over the use of data' (*Garantie der Datennutzungskontrolle*).[64]

Germany wants to create a data infrastructure that would guarantee European control over the data of European citizens and reduce dependence on foreign CSPs.[65] This is to be done by linking centralised and decentralised infrastructures (cloud and

59 Bundesregierung, *Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Andrej Hunko, Heike Hänsel, Ulla Jelpke, weiterer Abgeordneter und der Fraktion DIE LINKE*, BT-Drs. 19/3392, 2.

60 The Hessian Commissioner's statements are not limited to Microsoft, but also apply to other US CSPs. Der Hessische Beauftragte für Datenschutz und Informationsfreiheit, *Stellungnahme des Hessischen Beauftragten für Datenschutz und Informationsfreiheit zum Einsatz von Microsoft Office 365 in hessischen Schulen*, 09 July 2019, https://datenschutz.hessen.de/pressemitteilungen/stellungnahme-des-hessischen-beauftragten-für-datenschutz-und [14.04.2020].

61 Bundesministerium für Wirtschaft und Energie, *Das Projekt GAIA-X: Eine vernetzte Dateninfrastruktur als Wiege eines vitalen, europäischen Ökosystems*, https://www.bmwi.de/Redaktion/DE/Publikationen/Digitale-Welt/das-projekt-gaia-x.pdf [14.04.2020].

62 Ibid, 6.

63 Ibid, 7.

64 Ibid, 15.

65 Ibid, 9.

edge services) into one coherent system, based on open technologies and providing interfaces for the facilitation of data exchange and use of applications.[66] Crucially, this is to be done based on existing and yet-to-be-built European services and infrastructure, thereby limiting the exposure to US law enforcement by cutting out US headquartered CSPs. It is not surprising that these US companies, feeling the threat of a loss of market share in the important European market, are intensively lobbying against such autarky, rather singing the praises of the benefits of cooperation and investing in data centres located in key European States to alleviate concerns about data localisation.[67]

D. Poland

However, not every European State follows the path of achieving digital sovereignty through the exclusion of US CSPs from access to key data. In 2018, the Polish government launched the programme 'Common Information Infrastructure of the State' (*Wspólna Infrastruktura Informatyczna Państwa*, WIIP), which aims at creating two public cloud services: Public Computational Clouds (*Publiczne Chmury Obliczeniowe*) and a Governmental Computational Cloud (*Rządowa Chmura Obliczeniowa*).[68] With this, the Polish government does not exclude foreign CSPs, but rather applies different security and access standards to different types of data. For instance, the Public Computational Cloud (or simply 'National Cloud', *Chmura Krajowa*) will be set up in partnership with Google, which will build a Google Cloud hub in Warsaw.[69]

Currently, the largest and most strategically important client of the National Cloud is the largest bank in Poland, PKO BP and the National Cloud is aimed predominately at the private sector. Public and local administration will be able to use the Governmental Computational Cloud, which is currently in the phase of planning and in November 2019 issued a call for expressions of interest by those entities that would like to take part in a tender for setting up such a cloud service.[70] For this public cloud, the government will set up security requirements and a Governmental Security Cluster (*Rządowy Klaster Bezpieczeństwa*), presumably for the most sensitive data.[71]

66 Ibid, 12.

67 See Sabine Bendiek, 'Digitale Souveränität durch Partnerschaft: Wie Deutschland und Europa ihre Cloud-Zukunft selbstbestimmt gestalten können', (*Microsoft Blog*, 28 October 2019), https://www.microsoft.com/de-de/berlin/artikel/digitale-souveraenitaet-durch-partnerschaft.aspx [14.04.2020].

68 Poland Ministry of Digital Affairs decision on Common Infrastructure of the State: Ministerstwo Cyfryzacji, *Wspólna Infrastruktura Informatyczna Państwa*, https://www.gov.pl/web/cyfryzacja/wspolna-infrastruktura-panstwa-wip-20 (last modified 27.09.2019 12:11) [14.04.2020].

69 Operator Chmury Krajowej, 'Strategiczne partnerstwo Operatora Chmury Krajoweji Google dla cyfryzacji polskiej gospodarki', (Press release, 27 September 2019), https://chmurakrajowa.pl/pdf/informacja_prasowa_27.09.2019.pdf [14.04.2020].

70 Michał Duszczyk, 'W 2020 roku państwo przeniesie się do chmury', (*Rzeczpospolita*, 05 October 2019), https://cyfrowa.rp.pl/it/41069-w-2020-roku-panstwo-przeniesie-sie-do-chmury [14.04.2020].

71 Ibid.

It remains to be seen whether Poland will exclude foreign CSPs from this Governmental Security Cluster or try to secure governmental data contractually and through encryption. It has to be noted, however, that Poland cannot rely on big national CSPs and therefore is dependent on outside expertise for its national cloud and is thus limited in a potential quest for digital sovereignty.

E. Preliminary Conclusions

By way of a preliminary conclusion, we can see that there is no one universal way in which States try to ensure their data sovereignty against the challenge posed by the US CLOUD Act and the potential long-arm jurisdiction over European data stored with US CSPs. The potential reactions range from data localisation laws such as in Russia or China,[72] through treaty-based cooperation with the United States (the US-UK Agreement, for example), contract-based cooperation with US CSPs using software solutions while building a localised data cluster (Poland), to escaping from US long-arm jurisdiction by legislative decoupling (France) and technological independence (Germany). All these examples, however, show that States recognise the need to ensure sovereignty over their own and their citizens' data and limit its exposure to the control and access by other sovereigns, in particular the US.

5. RECOMMENDATIONS: BALANCING SOVEREIGNTY

The preceding analysis has shown that Western States in general and European States in particular are increasingly conscious of challenges to their sovereignty, understood as the capability for autonomous action, that stem from the rapid development of digital technologies. Especially with regard to the rising importance of personal, business and governmental data in digital (data-driven) economies and in view of US technological dominance in the sector of cloud storage, cloud services and data processing, these States frame their sovereignty in terms of exclusive control over data stored in the cloud, to the exclusion of third States acting through their organs, for instance law enforcement agencies. Therefore, it becomes a priority to find solutions which reconcile the continued consumption of services of foreign-headquartered CSPs, the needs of law enforcement and the protection of sensitive data from unauthorised or excessive access by law enforcement agencies of third States.

The preceding analysis also discusses different ways how States such as the UK, France, Germany or Poland address these issues of data sovereignty vis-à-vis US CSPs in view of the US CLOUD Act's long-arm jurisdiction over foreign data stored by these CSPs. In our view, the concerns raised by France and Germany over their data sovereignty are not confined to those States, but describe a universal challenge to and evolution of the understanding of the principle of sovereignty in cyberspace.

72 John Selby, 'Data Localization Laws: Trade Barriers or Legitimate Responses to Cybersecurity Risks, or Both?', (2017) International Journal of Law and Information Technology, Volume 25, Issue 3, 213–232.

States will increasingly face difficult policy decisions with regard to deciding how best to balance competing sovereign interests. Based on the described policy and legal approaches to 'data sovereignty', we propose seven actions for consideration by States that have not yet specifically addressed the issues discussed in this paper. Actions A, B, C and D are likely to be completed sequentially, followed by actions E, F and G, which could be executed concurrently. Of course, all of these recommendations involve speculation by the authors and will require further debate.

A. Formulate a Domestic Policy for Cloud Storage and Take a Position on 'Data Sovereignty'

Addressing the described challenges of control over data in cyberspace requires a two-step analytical exercise. *Firstly*, the government should formulate a domestic policy for cloud storage, taking into account the problems described above. *Secondly*, the government should analyse the issues regarding the interpretation of the principle of sovereignty and its application in cyberspace, especially with regard to the questions of jurisdiction and data sovereignty. This position should not only address the question of the applicability of international law to cyber operations, as done by various other governments, but also the challenges of jurisdiction, in particular with regard to data, as the Dutch government has done.[73] Such a position is especially important for States which do not have sovereign CSPs or where many companies and government departments are consuming cloud facilities from CSPs headquartered in a foreign State.

B. Enact Rules for the Distribution or Sharing of Sensitive Data

Pass a bill (and enact into law) rendering unlawful the distribution or sharing of the country's sovereign State data without permission from the government, including that which is stored in cloud capacities in other nations. Much like France's legislative[74] and industrial policy to protect French data against the reach of foreign States (including the US) and Germany's Hessian Commissioner's decision to forbid schools in the Land of Hesse to store students' personal data in the cloud if the providers are subject to US law,[75] other States must also move to manage the potential risk to their government and citizens if such data were to escape the State's control.

73 Dutch Ministry of Foreign Affairs, *Letter of 5 July 2019 from the Minister of Foreign Affairs to the President of the House of Representatives on the international legal order in cyberspace, Appendix*, (Government document, 5 July 2019), https://www.government.nl/ministries/ministry-of-foreign-affairs/documents/parliamentary-documents/2019/09/26/letter-to-the-parliament-on-the-international-legal-order-in-cyberspace [14.04.2020].

74 Loi n° 68-678 du 26 juillet 1968 relative à la communication de documents et renseignements d'ordre économique, commercial, industriel, financier ou technique à des personnes physiques ou morales étrangères', https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000501326&categorieLien=cid [14.04.2020].

75 See Der Hessische Beauftragte für Datenschutz und Informationsfreiheit, *Stellungnahme des Hessischen Beauftragten für Datenschutz und Informationsfreiheit zum Einsatz von Microsoft Office 365 in hessischen Schulen*, 09 July 2019, https://datenschutz.hessen.de/pressemitteilungen/stellungnahme-des-hessischen-beauftragten-für-datenschutz-und [14.04.2020].

Possible positions include that the State's data can be accessed only with permission, thereby obliging nations to notify when such data is delivered to or requested by another nation. This action may seem particularly challenging as it requires political support; however, the politicians of the particular nation should be made aware of the challenges their nation faces and that there is a roadmap to make them manageable.

C. Clearly Classify Data

To represent data that is to remain within the State's borders, some States mark the content with limited dissemination control marking.[76] States should consider a similar marking or use of an existing marking that represents data sovereignty to identify information that may reside in a foreign country or not.

Besides technical requirements to transmit or store data abroad which remain the primary focus for some countries, States should consider defining legal requirements for storing data abroad. As an example, some Canadian medical clinics operated as corporate entities store their patient data in the cloud (or in software that is backed up in the cloud), without fully understanding the risks related to such decisions. The legal requirements for storing citizen or government data in the cloud should also extend to corporations and similar entities.

D. Enter into Bilateral Agreements with the US, UK and EU

Negotiate and sign bilateral agreements, where applicable, on legitimate law enforcement access to data stored abroad with States that have adopted a national regulatory framework for cloud computing, such as the US (US CLOUD Act) and the UK (Crime Overseas Production Orders COPO Act 2019[77]). The purpose of such agreements would be to define, based on reciprocity, the scope of legitimate law enforcement access to data of the nation's legal and natural persons stored in data centres on the territory of other States and controlled by those other States' CSPs (e.g. the US, thus putting them within reach of the US CLOUD Act). However, the governments should also consider concluding separate agreements or including special provisions for governmental data, according to its inviolability for law enforcement purposes, similar to the Agreement between Estonia and Luxembourg for the hosting of data and information systems.[78]

76 Such as NOFORN to represent "no foreign dissemination", or CEO to represent "Canadian eyes only," in Canada. Admittedly, CEO marked assets and information are sometimes shared with "Foreign Integrees" who sign non-disclosure agreements. The GC is aware of the challenges this may bring and has requested all departments restrict CEO assets and information access to Canadians only. Canadian Committee On National Security Systems, *CCNSS Bulletin Edition 1*, March 2018, 3, https://www.cyber.gc.ca/sites/default/files/publications/ccssn-1-eng.pdf [14.04.2020].

77 Crime (Overseas Production Orders) Act 2019, c. 5.

78 *Agreement between the Grand Duchy of Luxembourg and the Republic of Estonia on the hosting of data and information of 20 June 2017, as appended to Loi du 1er décembre 2017 portant approbation du « Agreement between the Grand Duchy of Luxembourg and the Republic of Estonia on the hosting of data and information systems »*, http://data.legilux.public.lu/eli/etat/leg/loi/2017/12/01/a1029/jo [14.04.2020].

It may seem too soon to begin negotiations with the EU with regard to the proposed EU e-evidence regulation (EU COM/2018/225 final); however, the US and the EU have already jointly announced that negotiations[79] are underway on the matter. Negotiating with the EU now is important for nations that, like Canada, have arrangements with providers headquartered in the EU.[80]

The agreements offer the opportunity to discuss the challenges and construct an agreement in line with each other's newly enacted laws and best interests.

E. Advise Departments of the Challenge

Considering that various departments are already storing their information in non-sovereign cloud facilities, it is important to advise and educate all departments on the challenges placed on the nation by their actions and the legal and political complications that may arise if a State's data, of both non-national interest and national interest, were to be accessible by another nation or breached.

Consider encouraging departments currently using non-sovereign cloud capacities to migrate their information to sovereign cloud capacities within a reasonable timeline.

F. Mandate International Interaction

Put in place a mandate to interact with other nations to better understand and be aware of their legal positions and changes to them. This action may involve various departments[81] to manage the discussions, digest the effect of international positions and disseminate the information to the rest of the government.

G. Cultivate Sovereign CSPs

Like Germany's GAIA-X project, States should consider creating a national programme to foster and promote nationally headquartered companies to invest in creating and offering CSP services within their country. These services could be coupled with cross-departmental agreement to host government-used services from within a government-owned and -operated data centre, thereby supporting national and non-national interests.

79 US Department of Justice, *Joint US-EU Statement on Electronic Evidence Sharing Negotiations*, 26 September 2019, https://www.justice.gov/opa/pr/joint-us-eu-statement-electronic-evidence-sharing-negotiations [14.04.2020].

80 The Shared Services Canada (SSC) department has arrangements with 8 providers; one of them, OVH, is headquartered in France while 6 others are headquartered in the US. See SSC's website to help understand its GC cloud broker responsibility, Shared Services Canada, *Cloud services*, 13 August 2019, https://www.canada.ca/en/shared-services/corporate/cloud-services.html [14.04.2020].

81 Likely examples for Canada: Global Affairs, Justice, Privy Council Office, and/or Treasury Board.

6. INSTEAD OF A CONCLUSION

This article has presented recommendations for States to consider the management of their sovereign data. The enactment of these recommendations could help governments to formulate a comprehensive data sovereignty strategy which balances the need to protect and retain control over sensitive data while at the same time being open to international cooperation in addressing the legitimate needs of law enforcement. The alternative – strict data localisation laws as seen in Russia[82] – might lead to the increasing fragmentation of cyberspace and endanger the goal of promoting an open, secure, stable, accessible and peaceful ICT environment, which the international community endorses.[83] We will see over the coming months and years how and whether more States choose to address the matter of data sovereignty and what their conclusions will be.

82 Federal Law No. 242-FZ of July 21, 2014 on Amending Some Legislative Acts of the Russian Federation in as Much as It Concerns Updating the Procedure for Personal Data Processing in Information-Telecommunication Networks (with Amendments and Additions, official translation available at: https://pd.rkn.gov.ru/authority/p146/p191/ [14.04.2020].

83 See UN General Assembly, *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, para. 24.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Legal Issues Related to Cyber Threat Information Sharing Among Private Entities for Critical Infrastructure Protection

Livinus Obiora Nweke
Information Security and
Communication Technology
Norwegian University of Science and
Technology (NTNU)
Gjøvik, Norway
livinus.nweke@ntnu.no

Stephen Wolthusen
School of Mathematics and
Information Security
Royal Holloway, University of London
Egham, United Kingdom
stephen.wolthusen@rhul.ac.uk
Information Security and
Communication Technology
Norwegian University of Science and
Technology (NTNU)
Gjøvik, Norway
stephen.wolthusen@ntnu.no

Abstract: The menace of cyber attacks has become a concern for both the public and private sectors. Several approaches have been proposed to tackle the challenge, but an approach that has received widespread acceptance among cyber security professionals in both public and private sectors is cyber threat information (CTI) sharing. CTI refers to any information that can help an organisation identify, assess, monitor and respond to cyber threats. It includes indicators of compromise; tactics, techniques and procedures used by threat actors; suggested actions to detect, contain, or prevent attacks; and the findings from the analyses of incidents. Sharing CTI has been proposed as an efficient and effective way of improving overall cyber intelligence and defence. However, there are sources of liability that may dissuade private entities from participating in such sharing. The most cited source of liability is privacy and data protection law; although antitrust law, tort of negligence law and intellectual property law are also

cited as potential sources of liability. In this study, we review the extent to which the provisions of privacy and data protection law support or refute the sharing of CTI. This will provide guidance and incentives for private entities willing to participate in CTI sharing, especially for critical infrastructure protection.

Keywords: *legal issues, CTI sharing, GDPR, critical infrastructure protection*

1. INTRODUCTION

In recent years, the cost of cyber incidents has been rising. The Internet Society's Online Trust Alliance (OTA) reports that more than 2 million cyber incidents occurred in 2018, resulting in over $45 billion in losses.[1] The report notes that the financial impact of ransomware rose by 60%, losses from business email compromise doubled and crypto-jacking incidents more than tripled. Attacks on critical infrastructure are also expected to rise. For instance, the Department of Homeland Security in the United States (US) observes that 54% in the utility sector expect a cyber attack on critical infrastructure in 2020.[2] Considering the complexities in the cyber threat landscape, organisations can no longer rely on internally generated cyber threat intelligence (CTI) to protect themselves against these rising threats. Thus, CTI sharing has been proposed as an efficient and effective way of improving overall cyber intelligence and defence.

CTI sharing involves exchanging information relating to threat intelligence between entities, usually of a similar nature, for the purpose of enhancing their security posture by exploiting their collective knowledge, experience and capabilities.[3] Several studies have shown that CTI sharing is an effective tool for organisations to protect themselves against cyber attacks.[4] It enables organisations to understand trending cyber attacks and to implement the most efficient and effective strategies in combating those attacks.

1 Internet Society's Online Trust Alliance (OTA), '2018 cyber incident & breach trends report' (OTA, 9 July 2019) <https://www.internetsociety.org/wp-content/uploads/2019/07/OTA-Incident-Breach-Trends-Report_2019.pdf> accessed 11 December 2019.

2 Homeland Security Today, '54 Percent in Utility Sector Expect Cyber Attack on Critical Infrastructure in Next Year' (Homeland Security Today, 8 October 2019) <https://www.hstoday.us/subject-matter-areas/infrastructure-security/54-percent-in-utility-sector-expect-cyber-attack-on-critical-infrastructure-in-next-year/> accessed 16 December 2019.

3 National Institute of Standards and Technology (NIST), *Guide to Cyber Threat Information Sharing* (NIST Special Publication 800-150 2016) iii.

4 Cristin Goodwin and J. Paul Nicholas, *A Framework for Cybersecurity Information Sharing and Risk Reduction* (Microsoft 2015) 3.

There are several contexts in which CTI can be shared. It can be from a government to another government or to private entities; private entities sharing CTI with each other; or when private entities share CTI in their possession with the government.[5] In this paper, we examine CTI sharing in the context of private entities sharing cyber intelligence with each other: for example, when several companies in a sector (for example, the critical infrastructure sector) establish a formal exchange or formal agreements to share relevant CTI.[6] Such sharing frameworks would enable private entities to leverage the shared knowledge and techniques to better protect their assets while assisting others to do the same.

Private entities that wish to share CTI in their possession with others are faced with legal questions and would have to consider if any information they intend to share contains material that is potentially protected under data protection and privacy law, antitrust law, tort of negligence law, or intellectual property law. We focus on data protection and privacy law as it has shown to be the source of greatest concern, discouraging private entities willing to participate in CTI sharing. We consider the provisions of laws and regulations in the European Union (EU), Norway and the US related to CTI sharing, as those in the US and EU are models for many jurisdictions around the world.

In this paper, we first present the basic concepts of CTI sharing, including the existing CTI sharing architectures, benefits and challenges. We then provide a survey of the existing laws and regulations, which will serve as the basis for providing guidance and incentives for private entities willing to participate in CTI sharing. Lastly, we present a discussion on how well the existing laws and regulations address the concerns of private entities that are willing to participate in CTI sharing with each other. By reviewing the extent to which the provisions of the laws and regulations support or refute the sharing of CTI, we hope to provide guidance and incentives for private entities willing to participate in CTI sharing, especially for critical infrastructure protection.

The rest of this paper is organised as follows. Section 2 presents basic CTI sharing concepts including the existing CTI sharing architectures, the benefits and the challenges. Section 3 provides a survey of laws and regulations in the EU, Norway and the US related to CTI sharing; it also discusses the current trends among practitioners related to the legal implications of CTI sharing among private entities. Section 4 presents a discussion of how well the existing laws and regulations address the concerns of private entities willing to participate in CTI sharing. Section 5 concludes the paper and suggests future work.

5 Andrew Nolan, *Cybersecurity and Information Sharing: Legal Challenges and Solutions* (Congressional Research Service 2015) 5.

6 Ibid. 6.

2. BACKGROUND

In this section, we present basic CTI sharing concepts including the existing CTI sharing architectures. We also explore the benefits and challenges to provide the necessary background for an understanding of the legal issues related to CTI sharing among private entities.

A. Existing CTI Sharing Architectures

CTI refers to any information that can help an organisation identify, assess, monitor and respond to cyber threats. It includes indicators of compromise; the tactics, techniques and procedures (TTPs) used by threat actors; suggested actions to detect, contain or prevent attacks; and the findings from the analysis of incidents.[7] It is no longer the case that organisations must rely only on internal threat intelligence for protection from ever-evolving cyber threats. Hence, the sharing of CTI between entities usually of a similar nature has been proposed as an efficient and effective approach for addressing the complexities of the cyber threat landscape.

Two basic CTI sharing architectures may be adopted by private entities willing to share CTI. The first approach is the use of a centralised architecture, where a central organisation is responsible for the exchange of CTI among the participating entities and may have to perform additional processing to enrich the information.[8] The central body ensures interoperability by using open, standard data formats and transport protocols to provide timely and seamless portability of CTI. Typical examples of centralised architecture are the Information Sharing and Analysis Centres (ISACs).

ISACs provide a central resource for collecting information on cyber threats (in many cases relating to critical infrastructure) and facilitate active sharing of information between the private and the public sectors.[9] They are usually trusted entities that are constituted by representatives of critical infrastructure owners and operators. ISACs were originally created in the US after the first terrorist attacks on the World Trade Centre. The main objective was to identify opportunities for cooperation between the public and private sectors for the protection of US critical infrastructure.[10] European legislation also advocates cooperation in cybersecurity which the creation of ISACs represents. For example, the NIS Directive encourages incident reporting and the sharing of information with computer security incident response teams (CSIRTs) which involves the sharing of threat intelligence.[11]

7 National Institute of Standards and Technology (NIST), *Guide to Cyber Threat Information Sharing* (NIST Special Publication 800-150 2016) ii.
8 Ibid. 17.
9 European Union Agency for Network and Information Security (ENISA), *ENISA'S Opinion Paper on ISAC Cooperation* (Opinion Paper 2019) 3.
10 Ibid. 3.
11 Ibid. 4.

The second CTI sharing architecture is the peer-to-peer architecture, where private entities that are willing to share CTI with each other do so directly without an intermediary. This type of architecture enables great agility in that participants can receive CTI directly from the source and the problem of having a single point of failure as in the case of centralised architecture is eliminated.[12] A typical example of a peer-to-peer architecture of CTI sharing can be found in the power sector.[13]

Regardless of which CTI sharing architecture an organisation decides to adopt, there is a need to establish information sharing rules before proceeding. The NIST guide to CTI sharing recommends the following rules:[14]

- List the types of threat information that may be shared.
- Describe the conditions and circumstances when sharing is permitted.
- Identify approved recipients of threat information.
- Describe any requirements for redacting or sanitising information to be shared.
- Specify if source attribution is permitted.
- Apply information handling designations that describe recipient obligations for protecting information.

These rules would help to ensure that the publication and dissemination of threat information are controlled. The goal is to prevent the sharing of information that, if not properly handled, may have serious legal implications for the organisation.[15]

However, these rules are not quite complete as far as NIST provides. Specifically, the issue of sanitising information is unfortunately not something that can be solved based on a single record. With multiple anonymised records or queries, it will be possible to de-anonymise or otherwise fill in the gaps of queries. So, one has to either accept that sanitising offers only a weak form of anonymity and prevention of leaking sensitive information or has to use far more restrictive measures.

B. Benefits of CTI Sharing

CTI sharing provides organisations with access to threat information that ordinarily they may not have been able to obtain without participating in such a sharing endeavour. Organisations can exploit these shared resources to improve their overall security

12 National Institute of Standards and Technology (NIST), *Guide to Cyber Threat Information Sharing* (NIST Special Publication 800-150 2016) 17.

13 Steve Livingston, Suzanna Sanborn, Andrew Slaughter and Paul Zonneveld, 'Managing Cyber Risk in the Electric Power Sector: Emerging Threats to Supply Chain and Industrial Control Systems' (Deloitte Insights, 2018) <https://www2.deloitte.com/content/dam/insights/us/articles/4921_Managing-cyber-risk-Electric-energy/DI_Managing-cyber-risk.pdf> accessed 11 April 2020.

14 National Institute of Standards and Technology (NIST), *Guide to Cyber Threat Information Sharing* (NIST Special Publication 800-150 2016) 10.

15 Ibid. 5.

posture by using the knowledge, experience and capabilities of the participating entities. This ensures that the detection of one organisation becomes the prevention of another.[16]

There are several ways that an organisation can use the shared threat information. It might use the information for operational purposes, such as updating its security controls for continuous monitoring with new indicators and configurations to detect the latest attacks and comprises.[17] The shared threat information might also be used strategically, such as when planning major changes to an organisation's security structure.[18]

Sharing CTI between entities of a similar nature can be greatly beneficial because participating entities will often face actors that use similar TTPs and target the same types of infrastructures. Defending against cyber threats is much more effective and efficient when organisations collaborate to defend against well-organised and capable actors.[19] This type of alliance will enable organisations to mitigate risks and ameliorate their overall security readiness.

The additional benefits of CTI sharing have been identified as including the following: shared situational awareness, where organisations exploit the collective knowledge, experience and analytical capabilities of the participating entities; improved security posture, which allows organisations to implement protective measures, improve detection capabilities and more effectively respond to and recover from incidents based on observed trends in the threat landscape; knowledge maturation, which enriches the value of threat information; and greater defensive agility, where participating entities adapt quickly to evolving threats.[20] Whilst there are benefits in CTI sharing, it still poses some challenges that need to be considered, some of which are explored in the following subsection.

C. Challenges of CTI Sharing

One of the prerequisites to CTI sharing involves establishing a trust relationship among the participating entities.[21] This process can be very challenging, as building trust requires a lot of work to develop and sustain it. However, an organisation's ability to establish trust between entities willing to share CTI is pivotal to the success of any CTI sharing scheme. Hence, the cost and effort required to build a trust relationship

16 Ibid. 3.
17 Cristin Goodwin and J. Paul Nicholas, *A framework for cybersecurity information sharing and risk reduction* (Microsoft 2015) 10.
18 National Institute of Standards and Technology (NIST), *Guide to Cyber Threat Information Sharing* (NIST Special Publication 800-150 2016) 3.
19 Ibid. 3.
20 Ibid. 3-4.
21 Cristin Goodwin and J. Paul Nicholas, *A framework for cybersecurity information sharing and risk reduction* (Microsoft 2015) 3.

among participating entities may discourage an organisation's willingness to join in such a sharing scheme.

Achieving interoperability and automation have also been cited as challenges to CTI sharing.[22] The problem of interoperability seems to be more profound for organisations that adopt peer-to-peer sharing architecture than for those that choose centralised architecture. However, both types of sharing architectures must deal with the additional complexities introduced by automation. With the use of automation, the participating entities would have to agree on the data format and methodology to be employed. All these require organisations to invest additional resources in ensuring that the shared CTI can be automated and be easily reusable by the participating entities.

Organisations participating in CTI sharing may not want to disclose their identity to avoid a perceived risk to the organisation's reputation. The unwillingness to disclose their identity could be problematic as the credibility of the shared threat information may be brought into disrepute.Also, it is natural for participating entities to doubt the credibility of shared information if its source is unknown. Therefore, organisations willing to participate in CTI sharing may have to weigh the perceived risk to the reputation of the organisation against the dangers of not sharing threat information.

Another challenge that may discourage private entities from participating is the problem of incomplete or false information. This means that there is the possibility of any of the participating parties sharing incomplete or false information which may contaminate or mislead the algorithms or analysts. In such a scenario, the danger is that it either disincentivises sharing or encourages other participating entities to share questionable information. Any liability waiver usually becomes void when negligence is involved, so there are some data quality obligations inherent in CTI sharing arrangements that must be considered.

Legal liability that may arise from CTI sharing is a major source of concern for organisations willing to participate in such sharing schemes.[23] This is because the legal issues relating to CTI sharing tend to be complex and they have very few certain resolutions.[24] Various laws and regulations have been proposed and implemented to address such concerns. For example, the Cybersecurity Information Sharing Act (CISA) was approved by the US Congress in 2015 to provide legal protection for organisations that participate in CTI sharing. In Europe, a similar cybersecurity framework offers the same protection against any liability that may result from CTI

22 National Institute of Standards and Technology (NIST), *Guide to Cyber Threat Information Sharing* (NIST Special Publication 800-150 2016) 4.

23 Andrew Nolan, *Cybersecurity and Information Sharing: Legal Challenges and Solutions* (Congressional Research Service 2015) 5.

24 Ibid.

sharing for the protection of network and information systems across the Union.[25] These legal protections require organisations to follow set rules when sharing CTI. The next section provides a review of these laws and regulations to assess the extent to which they support or refute the sharing of CTI among private entities.

3. LAWS AND REGULATIONS RELATED TO CTI SHARING

Various laws and regulations have been proposed to encourage CTI sharing and we provide a survey of these laws and regulations in this section. The purpose of this review is to assess their provisions, which will then serve as the basis for providing guidance and incentives for private entities willing to engage in CTI sharing.

A. Laws and Regulations in the European Union (EU)

A good number of laws and regulations have been proposed in the EU over the years to promote the sharing of CTI. The most relevant of these are Directive (EU) 2016/1148 of 6 July 2016,[26] also known as the network and information systems (NIS) Directive; the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679 of 27 April 2016);[27] and the EU Cybersecurity Act (Regulation (EU) 2019/881 of 17 April 2019).[28] In the EU, a Regulation is a binding legislative act that is directly applicable in its entirety across the EU; while a Directive is a legislative act that stipulates goals that all EU countries must achieve (minimum-level legal provisions), but it is incumbent on the individual countries to promulgate their own laws in order to reach these goals.[29]

The NIS Directive can be considered the first EU-wide cybersecurity legislation. It aims to enhance cybersecurity across the EU. The directive encourages the sharing of CTI for the protection of critical infrastructure by providing an enabling environment for setting up ISACs which will foster the sharing CTI within and between the EU member states. Following the adoption of the NIS directive in 2016, it became an EU

25 Dimitra Markopoulou, Vagelis Papakonstantinou and Paul de Hert, 'The new EU cybersecurity framework: The NIS Directive, ENISA's role and the General Data Protection Regulation' (2019) 35(6) Computer Law and Security Review <https://www.sciencedirect.com/science/article/pii/S0267364919300512> accessed 12 April 2020.

26 Directive (EU) 2016/1148 concerning measures for a high common level of security of network and information systems across the Union [2016] OJ L194/1.

27 Regulation (EU) 2016/679 on the protection of natural persons with regards to the processing of personal data and on free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1.

28 Regulation (EU) 2019/881 on ENISA (the European Union Agency for Cybersecurity) and on information and communication technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act) [2019] OJ L151/15.

29 European Union, 'Regulations, Directives and other acts' (EU Law, 7 March 2019) < https://europa.eu/european-union/eu-law/legal-acts_en> accessed 20 December 2019.

Directive requiring that every member state adopt national legislation which follows or 'transposes' the directive.[30] In general, the NIS Directive has three main parts:[31]

- **National capabilities:** EU member states must have certain national cybersecurity capabilities such as a national CSIRT and must perform cyber exercises, etc.
- **Cross-border collaboration:** Cross-border collaboration between EU countries, including the operational EU CSIRT network and the strategic NIS cooperation group.
- **National supervision of critical sectors:** EU member states must supervise the cybersecurity of critical market operators in their country: ex-ante supervision in critical sectors (energy, transport, water, health and finance), ex-post supervision for critical digital service providers (internet exchange points, domain name systems, etc).

The NIS Directive observes that the 'responsibilities in ensuring the security of network and information systems lie, to a great extent, with operators of essential services'.[32] It does differentiate between sectors, placing higher burdens on critical infrastructure operators. The implication of this is that private entities that provide essential services (critical infrastructure operators) are obliged to ensure the protection of their network and information systems. The NIS Directive encourages a culture of risk management, which include risk assessment and the implementation of appropriate security measures for the protection of network and information systems within the critical infrastructure sector. Among these measures is the sharing of CTI.

Regulation (EU) 2016/679,[33] or GDPR as it is better known, has been hailed as the model for data protection and privacy laws both in Europe and beyond.[34] The goal of the Regulation is to harmonise data and privacy laws across Europe, to increase the levels of protection for EU citizens and to give them greater control over their personal data. The regulation 'protects fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data'.[35] It has also redefined the way organisations across Europe and how those who offer goods and/or services to EU citizens around the globe, process personal data. GDPR contains provisions and requirements that are related to the processing of personal data of individuals

30 Directive (EU) 2016/1148 concerning measures for a high common level of security of network and information systems across the Union [2016] OJ L194/1.
31 Ibid.
32 Ibid.
33 Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1.
34 Clare Sillivan and Eric Burger, '"In the public interest": The privacy implications of international business-to-business sharing of cyber-threat intelligence' (2017) 33(1) Computer Law and Security <https://www.sciencedirect.com/science/article/pii/S0267364916302229> accessed 21 December 2019.
35 Regulation (EU) 2016/679 Art 1.

(data subjects) inside the European Economic Area (EEA). These provisions and requirements include the provisions that cover the scope, application and objectives of the data protection regulations and the implementing arrangements.

The EU Cybersecurity Act's main objective is to provide a permanent mandate for the ENISA and to establish a cybersecurity certification framework. It strengthens ENISA through the provision of more resources and a legal framework to improve cybersecurity capabilities at Union level, among member states, Union institutions, bodies, offices and agencies and relevant private and public stakeholders on matters related to cybersecurity.[36] Among the provisions of the EU Cybersecurity Act, the provision that is most relevant to this study is Article 6(2), which states that 'ENISA shall support information sharing in and between sectors, in particular in the sectors listed in Annex II to Directive (EU) 2016/1148, by providing best practices and guidance on available tools and procedures, as well as on how to address regulatory issues related to information-sharing'.[37]

B. Laws and Regulations in Norway and the US

In this subsection, we examine the laws and regulations in Norway and the US to review efforts in other countries outside the EU regarding CTI sharing. Norway is a member of the European Economic Area (EEA) and so some EU regulations are also applicable. Like other EEA member states, Norway is required to promulgate laws in line with EU Directives if they are relevant to the EEA. The Norwegian National Security Act (Security Act) is the most relevant law in Norway to this study. In the US, the Cybersecurity Information Sharing Act of 2015 (CISA) is considered to be the most significant cyber-related legislation as it establishes a mechanism for cybersecurity information sharing among private sector and government entities.[38] CISA has greatly impacted the sharing of CTI not just in the US but also around the world; thus, deserves consideration.

The Security Act took effect on January 1, 2019. Its purpose is threefold: to safeguard Norway's sovereignty, territorial integrity and democratic governance and other national security interests; to prevent, detect and counteract security threats; and to ensure that security measures are implemented in accordance with basic legal principles and values in a democratic society.[39] It is mainly concerned with security-rated information, information systems and objects or infrastructure essential for basic national functions (critical infrastructure). It applies to state, county and municipal bodies and to suppliers of goods or services that can access or produce security-classified information.[40] For example, Article 2(3) requires that 'the security

36 Ibid.
37 Ibid. Art 6.
38 John Heidenreich, 'The Privacy Issues Presented by the Cybersecurity Information Sharing Act' (2015) 91(395) North Dakota Law Review < https://law.und.edu/_files/docs/ndlr/pdf/issues/91/2/91ndlr395.pdf> accessed 21 December 2019.
39 National Security Act (Norway) LOV-2018-06-01-24 (Security Act) [2018] Jnr 2018-0165 ch 1, art 1.
40 Ibid. ch 1, art 2-3.

authority shall ensure that businesses to which the law applies will have access to information on threat assessments and other information that is important for the companies' preventive security work'.[41] This implies that the Act not only supports an organisation's monitoring of its information systems to prevent, detect and counteract cyber incidents, it also offers greater flexibility to organisations when implementing such security measures including CTI sharing.

CISA was signed into law on December 18, 2015. The law has two main components: it authorises companies to monitor and implement defensive measures on their own information systems to counter cyber threats and it provides certain protections to encourage companies to share CTI.[42] Title I of the law is of greatest interest to private sector bodies willing to participate in cyber threat intelligence sharing. It states that 'non-federal entities can share CTI among themselves and with federal departments and agencies'.[43] It provides several safeguards which include protection from liability, non-waiver privilege and protection from Freedom of Information Act (FOIA) disclosure. Organisations that are covered by these protections must comply with CISA's requirements when participating in CTI sharing.

C. Legal Implications of CTI Sharing

We have provided a survey of the existing laws and regulations in the EU, Norway and the US related to CTI sharing. Our focus in this paper is on provisions that are related to personal data protection. A general theme of these laws and regulations is that CTI sharing is lawful but that care should be taken not to share information protected by data protection and privacy laws. In addition to the survey presented in the preceding section, we provide a discussion on the current trends among practitioners related to the legal implications of CTI sharing among private entities in this subsection.

Many authors have considered the extent to which the provisions of GDPR affect CTI sharing. Article 4(1) defines personal data as:

> 'any information relating to an identified or identifiable natural person ('data subject'). An identifiable natural person is one who can be identified, directly or indirectly, by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person'.[44]

41 Ibid. ch 2, art 3.

42 S.754 An Act to improve cybersecurity in the United States through enhanced sharing of information about cybersecurity threats, and for other purposes (Cybersecurity Information Sharing Act of 2015) [2015].

43 Ibid.

44 Regulation (EU) 2016/679 on the protection of natural persons with regards to the processing of personal data and on free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1, art 4.

CTI is likely to contain sensitive and identifying information such as IP and email addresses.[45] This may raise concerns for private entities willing to participate in CTI sharing as they must ensure conformance with legal and regulatory requirements.

Borden et al. have argued that CTI sharing is lawful under GDPR.[46] They observe that the provision of Article 6, which requires 'legitimate interests' for the processing of personal data in CTI, is satisfied by private entities participating in such a scheme. They also suggest that GPDR Recitals 47, 49 and 50 supports the processing of personal data for fraud prevention, ensuring network and information security and indicating possible acts or threats to public security. These are all goals of CTI sharing.

Sullivan and Burger discuss the legal issues related to international business-to-business sharing of cyber threat intelligence.[47] They opine that data protection and privacy laws affect the willingness of private entities to participate in CTI sharing. They use GDPR as a case study (considering that its requirements do not only apply to companies incorporated in the EU but also to third countries and international organisations) to investigate whether automated sharing of information between businesses may be legal. The study concludes that the sharing of cyber threat intelligence between businesses is likely to be necessary for the legitimate interests of the data controller under Article 6(1)(f) of GDPR and may be clearly justified and lawful on public interest grounds.

Similarly, Maltzan observes that Article 6(1)(f) of GDPR may be used as a legal ground for the processing of personal data when private entities participate in sharing of CTI with each other.[48] She maintains in the paper that the legitimate interest clause may allow the data controller to process personal data if none of the other circumstances listed in Article 6 of GDPR will suffice as a legal basis. She also notes that the lawfulness of CTI sharing under the provision requires an assessment of the test for validity based on the legitimacy and necessity of the processing and balance between the interests of the data controller and data subject. According to the Article 29 Working Party, 'this balance of interest test should consider issues of proportionality,

45 Adham Albakri, Eerke Boiten and Rogério De Lemos, 'Risks of Sharing Cyber Incident Information' In Proceedings of International Conference on Availability, Reliability and Security, Hamburg, Germany, August 27–30 2018 (ARES 2018) <https://dl.acm.org/doi/pdf/10.1145/3230833.3233284> accessed 18 December 2019.

46 Richard Borden, Joshua Mooney, Mark Taylor, and Matthew Sharkey, 'Threat Information Sharing Under GDPR' (American Bar Association, 6 March 2019) <https://www.americanbar.org/groups/science_technology/publications/scitech_lawyer/2019/spring/threat-information-sharing-under-gdpr/> accessed 20 December 2019.

47 Clare Sullivan and Eric Burger, ''In the public interest': The privacy implications of international business-to-business sharing of cyber-threat intelligence' (2017) 33(1) Computer Law and Security <https://www.sciencedirect.com/science/article/pii/S0267364916302229> accessed 21 December 2019.

48 Stephanie Von Maltzan, 'No contradiction between cyber-security and data protection? designing a data protection compliant incident response system' (2019) 10(1) EJLT <http://ejlt.org/article/view/665/893> accessed 22 December 2019.

the relevance of the personal data to the litigation and the consequences for the data subject'.[49]

Although our focus in this paper is on provisions related to personal data protection, other concerns may discourage private entities from participating in CTI sharing. Private entities that wish to share CTI may also have to consider if any of the information they intend to share contains material that is potentially protected under antitrust law, tort of negligence law or intellectual property law.[50] The laws and regulations that we have reviewed in this paper protect from liability for private entities only as long as they conform with the laid down requirements when sharing CTI, including removal of personal data that may be found in it. For example, the US Department of Justice released a statement clearly noting that CTI sharing does not raise antitrust issues.[51] It observes that private entities that participate in such sharing activities do not violate antitrust laws as the shared information is very technical in nature and very different from the sharing of competitively sensitive information such as current or future prices and output or business plan.

In general, the greatest concern for private entities willing to participate in CTI sharing is to consider whether any of the information they intend to share contains material that is protected by data protection and privacy laws. However, processing of CTI and subsequent sharing with others for the protection of network infrastructure can be viewed as 'legitimate interests'. Therefore, in agreement with the studies discussed above, we note that Article 6(1)(f) of GDPR may be used as the legal basis for private entities to participate in sharing CTI and that the principles stated in Article 5 of GDPR still need to be observed.

4. DISCUSSION

In this section, we present a discussion on how well the existing laws and regulations address the concerns of private entities willing to participate in CTI sharing with each other. Ambiguity in laws and regulations often breeds litigation and the costs of litigation may be significant enough to deter private entities from engaging in CTI sharing. This section considers whether there are legal and regulatory requirements that make the identified concerns difficult to address.

49 Article 29 Data Protection Working Party WP 136 Opinion 4/2007 on the concept of personal data [2007] 01248/07/EN.

50 Andrew Nolan, Cybersecurity and Information Sharing: Legal Challenges and Solutions (Congressional Research Service 2015) 12.

51 Department of Justice and Federal Trade Commission, Antitrust Policy Statement on Sharing of Cybersecurity Information (Policy Statement, United States Department of Justice and Federal Trade Commission) [2014].

There is a consensus among the existing laws and regulations and the current discussion among practitioners that cyber threat sharing can be performed lawfully. However, organisations that wish to participate in CTI sharing among themselves would have to consider issues that could arise from the disclosure of personal information, breaches of contractual terms and disclosure of sensitive or classified information. For example, CISA offers several safeguards for private entities that participate in CTI sharing, which include protections from liability, non-waiver privilege and protections from FOIA disclosure.[52] These protections are likely to become void when negligence leads to the disclosure of personal information, breaches of contractual terms or disclosure of classified information.

Organisations must take care when sharing CTI containing personal information. However, when such sharing becomes necessary, Article 6(1)(f) of GDPR may serve as a legal basis. CTI containing personal data also raises additional concerns for automating the CTI sharing process. This requires private entities to invest additional resources. They may also have to consider the likelihood of the shared information containing personal information. Articles 25 and 32 of GDPR offer suggestions on how to implement technical and organisational measures to mitigate the risks associated with processing such data.[53] Organisations may have to examine how these technical and organisational measures can be included when deploying an automated CTI sharing system.

Another issue likely to make the legal and regulatory requirements difficult to address is the civil liability that may arise from breaches of contractual terms. For example, if a company were to give its trade secrets as part of a CTI exchange, this might expose its directors to civil liability. The disclosure of sensitive or classified information could make the legal and regulatory requirements that cause the identified concerns difficult to address, because such information may cause serious injury to the national interest.

It would also be interesting to investigate how the decision-making process can be supported in private entities. This will enable them to share CTI in compliance with existing laws and regulations. Albakri, Boiten and Lemos have presented a model for evaluating the legal requirements for supporting decision-making when sharing CTI in the context of GDPR.[54] They describe the effect that GDPR legal aspects may have on the sharing of CTI and have translated the existing legal provisions into

52 S.754 An Act to improve cybersecurity in the United States through enhanced sharing of information about cybersecurity threats, and for other purposes (Cybersecurity Information Sharing Act of 2015) [2015].

53 Regulation (EU) 2016/679 on the protection of natural persons with regards to the processing of personal data and on free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1.

54 Adham Albakri, Eerke Boiten and Rogério De Lemos, 'Sharing Cyber Threat Intelligence Under the General Data Protection Regulation' In: Naldi, M., Italiano, G.F., Rannenberg, K., Medina, M., Bourka, A. (eds.) *Privacy Technologies and Policy - 7th Annual Privacy Forum*, APF 2019, Rome, Italy, June 13-14, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11498, pp. 28–41. Springer (2019) <https://link.springer.com/chapter/10.1007/978-3-030-21752-5_3> accessed 19 December 2019.

rules to enable organisations to share CTI whilst being legally compliant with the requirements for sharing personal information.

However, the work by Albakri et al. can be extended to provide a holistic approach that can guide private entities willing to participate in CTI sharing.[55] The holistic approach for developing such a reference framework would involve extracting the legal requirements from the existing laws and regulations, in addition to the functional and non-functional requirements coming from the CTI sharing architectures. These requirements could then be translated into rules that would guide organisations when they share CTI. This type of framework would allow organisations to demonstrate that they satisfy the legal requirements for CTI sharing and encourage private entities to join such a scheme.

5. CONCLUSIONS

There is no doubt that CTI sharing increases the overall cyber intelligence and defence of organisations. We have conducted a review of existing laws and regulations in the EU, Norway and the US related to CTI sharing. First, we presented the basic concepts of CTI sharing including the existing CTI sharing architectures. We then explored the benefits and challenges of such sharing. We have observed that several laws and regulations have been proposed to encourage CTI sharing among private entities. However, private entities still cite data protection and privacy laws as the greatest concern, discouraging them from participating in CTI sharing.

Our study indicates that the processing of CTI and subsequent sharing with others in a bid to protect network infrastructure and improve overall cyber intelligence and defence can be considered 'legitimate interests' under GDPR for processing of any personal data that may be found in CTI. If none of the other circumstances listed in Article 6 can be invoked as a legal basis, the legitimate interest clause can suffice. Hence, Article 6(1)(f) of GDPR may serve as the legal basis for private entities to participate in CTI sharing, especially for critical infrastructure protection.

Future work will be directed towards considering approaches which organisations can employ to automate the CTI sharing process, and which will still conform with the requirements of existing laws and regulations. For example, Articles 25 and 32 of GDPR offer suggestions on how to implement technical and organisational measures

55 Adham Albakri, Eerke Boiten and Rogério De Lemos, 'Sharing Cyber Threat Intelligence Under the General Data Protection Regulation' In: Naldi, M., Italiano, G.F., Rannenberg, K., Medina, M., Bourka, A. (eds.) *Privacy Technologies and Policy - 7th Annual Privacy Forum*, APF 2019, Rome, Italy, June 13-14, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11498, pp. 28–41. Springer (2019) <https://link.springer.com/chapter/10.1007/978-3-030-21752-5_3> accessed 19 December 2019.

to mitigate the risks associated with the processing of personal data.[56] Thus, it is possible to evaluate these legal requirements for automating CTI sharing to translate the existing legal provisions into rules that will enable organisations to share CTI whilst being legally compliant.

56 Regulation (EU) 2016/679 on the protection of natural persons with regards to the processing of personal data and on free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1, art 25, 32.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Making the Cyber Mercenary – Autonomous Weapons Systems and Common Article 1 of the Geneva Conventions

Aleksi Kajander
MA Candidate
Tallinn Law School
Tallinn University of Technology
Tallinn, Estonia
aleksi.kajander@gmail.com

Agnes Kasper
PhD, Senior Lecturer
Tallinn Law School
Tallinn University of Technology
Tallinn, Estonia
agnes.kasper@taltech.ee

Evhen Tsybulenko
PhD, Senior Lecturer
Tallinn Law School
Tallinn University of Technology
Tallinn, Estonia
evhen.tsybulenko@taltech.ee

Abstract: Common Article 1 of the Geneva Conventions requires that states 'respect and ensure respect for' the Geneva Conventions 'in all circumstances'. In the new 2016 Commentary to the Convention, the existence of not only a negative obligation, but also a positive obligation of third countries to a conflict to prevent violations was confirmed. Hence, third countries must do everything 'reasonably in their power to prevent and bring such violations to an end'.

The use of autonomous weapons systems (AWS) is imminent in the future, as demonstrated by the Pentagon committing to spend $2 billion on research, with similar research programmes taking place in other countries. The buying and selling of these AWS is an equally impending part of the future. Consequently, inevitably a state that

is buying or being supplied with AWS will use them in a conflict. Therefore, suppliers of such systems will have to comply with the aforementioned positive obligation.

This paper will examine the positive obligation's impact on the state supplying AWS to a conflict. This includes the question of whether it will be their responsibility at the manufacturing stage to ensure that the system cannot violate the Geneva Conventions and – because autonomous systems are somewhat uncontrollable and unpredictable as they will also learn rather than only carrying out pre-programmed commands – whether the supplying state will be obligated to maintain a permanent tether to the supplied AWS to monitor them. The implications of tethering the supplied AWS may go well beyond ensuring compliance with international humanitarian law (IHL), and may include multiplying the leverage of the supplying state by turning the systems into 'cyber mercenaries'.

Keywords: *autonomous weapons, Geneva Convention, international humanitarian law, IHL*

1. INTRODUCTION

The development of autonomous technology is raising questions and shifting paradigms in a variety of fields such as transport, business and even governance. The military is no exception to this trend, as the possibilities for the military uses of autonomous technology are becoming increasingly apparent. However, as in other fields, the existing framework of laws was not created with autonomous systems in mind, and therefore its application to such systems is unclear. In the case of the military application of autonomous weapons systems (AWS), the application of the existing rules is literally a matter of life and death.

The Geneva Conventions have long been a cornerstone of international humanitarian law (IHL), and their application and interpretation have had fundamental effects on conflicts since their introduction.[1] They are now having to be examined in a new light, which creates new legal questions about their application.

An updated Commentary was released on the First Geneva Convention in 2016, which confirmed the existence of a positive external obligation under Common Article 1,

1 Lindsey Cameron, Bruno Demeyere, Jean-Marie Henckaerts, Eve La Haye, Heike Niebergall-Lackner, 'The updated Commentary on the First Geneva Convention – a new tool for generating respect for international humanitarian law' (2015), ICRC 97,1210.

whereby the High Contracting Parties 'undertake to respect and ensure respect for' the Convention in 'all circumstances'.[2] This positive obligation requires that the High Contracting Parties do 'everything reasonably in their power to prevent and bring such violations to an end'.[3]

This positive external obligation reaches a whole new dimension with the introduction of AWS, as a contracting party supplying them could potentially have unprecedented control over their supplied systems, whether by their programming or by the presence of a 'backdoor' enabling remote control. Either would significantly improve their ability to prevent IHL violations. However, the latter type of tethering, if required by Common Article 1, could also bring a new dimension to cyber warfare and have unintended military and political effects. Therefore, backdoors are a double-edged sword in the sense that, while they may bring added compliance, they will bring additional risk factors in the form of unintended third parties gaining access to the AWS.

Therefore, defining the parameters of this positive external obligation will be of utmost importance for states supplying such AWS, as it will impact both the design of those systems and the circumstances in which they can be supplied. This paper aims to analyse the relationship and implications of the positive external obligation in Common Article 1 concerning AWS and the states supplying them, particularly whether the supplying state is obliged to maintain a tether to the supplied systems.

2. COMMON ARTICLE 1

At its core, Common Article 1 (CA1) has a two-fold structure, the first part of which is to restate the principle of *pacta sunt servanda*: the binding nature of the treaty and the obligation of the parties to perform the treaty obligations in good faith.[4] This first obligation is evidenced by the wording of the Article, under which all High Contracting Parties (HCPs) 'undertake to respect' the convention in all circumstances. The first obligation is therefore relatively straightforward: to ensure that each party performs their obligations in good faith and respects the Conventions and the entire body of international humanitarian law binding upon that state. The reference to 'all circumstances' clarifies that the obligations of CA1 are always applicable both in peace and in more exceptional circumstances, a view confirmed by the 2016 Commentary.[5]

2 Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field (First Geneva Convention), 12 August 1949, 75 UNTS 31, Article 1.

3 International Review of the Red Cross, *Commentary on the First Geneva Convention: Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field, 2nd edition* [154].

4 Ibid. 143.

5 Ibid. 185.

However, the second obligation is far more ambiguous, as arguably there are many ways of 'ensuring respect', and moreover, the scope of this obligation may include an external dimension regarding the compliance of other states. Hence, the second obligation, to 'ensure respect' for the Convention in all circumstances, would go beyond the ordinary principle of *pacta sunt servanda* in the sense that the parties are not only obliged to perform their obligation in good faith, but also to ensure that others do so as well.[6] This second obligation derives from the addition of the words 'and to ensure respect' for the Convention, which, read in combination with the first obligation, could conceivably be directed outwards.

There is debate regarding the scope of the obligation to 'ensure respect', whether it is narrow and not directed towards other parties or broad and external as the updated 2016 ICRC Commentary states.[7] In essence, at the time of its adoption the obligation to 'ensure respect' was not considered to be external in nature, as evidenced by the *travaux préparatoires*.[8] However, those in favour of a broad scope argue that, since its adoption, the meaning of the provision has evolved through subsequent practice to include an external dimension.[9] The counterarguments point to the existing contrary state practice and the high standard of Article 31(3)(b) of the Vienna Convention on the Law of Treaties, which in their view requires that all parties accept or acquiesce to the subsequent practice for it to be relevant.[10]

Under the narrow view, the obligation of states to ensure respect contained in CA1 pertains only to their organs and those acting under their effective control.[11] This has severe implications regarding AWS, as without the external dimension of the broad scope it would be sufficient for HCPs to ensure that their AWS respect the Convention. This obligation would nevertheless extend to supplied AWS in the sense that they should not of their own accord encourage IHL violations under CA1.[12] However, should their supplied AWS be misused, CA1 would not provide an obligation to ensure compliance by the systems, as the supplying state does not have effective control over them. Consequently, under the narrow scope the supplying states would only have to ensure that their own AWS and any AWS they have effective control over respect the Convention, and that those supplied do not encourage violations.

6 Ibid. 154.
7 Theo Boutruche, Marco Sassoli, 'Expert Opinion on Third States' Obligation vis-à-vis IHL Violations under International Law, with a special focus on Common Article 1 to the 1949 Geneva Conventions' < https://www.nrc.no/resources/legal-opinions/third-states-obligations-vis-a-vis-ihl-violations-under-international-law/> accessed 21 February 2020.
8 Andrea Breslin, 'Reflections on the Legal Obligation to Ensure Respect' (2017), Journal of Conflict and Security law 22(1), 11.
9 Boutruche, Sassoli (nr 7) 7-8.
10 Tomasz Zych, 'The Scope of the Obligation to Respect and to Ensure Respect for International Humanitarian law', (2009) Windsor Yearbook of Access to Justice 27, 256.
11 Ibid. 270.
12 Ibid. 265.

It ought to be highlighted that should a tether enabling effective control of a supplied AWS exist, then arguably it will be within the scope of the obligation to 'ensure respect' for CA1 for the supplying state, even under the narrow view. However, the narrow view cannot require a supplying state to tether supplied AWS in the first place, as there is no obligation towards ensuring respect in regard to other states. Therefore, the design decision of whether supplied AWS are tethered will determine whether the CA1 obligation will apply after they are exported. Thus, regardless of which interpretation prevails, CA1's obligation to ensure respect will conceivably affect the design of AWS, for if a tether is included, then the supplying state must comply with that obligation even after the system has been supplied.

For the purposes of this paper, the obligation of 'ensuring respect' shall be construed to include an external dimension under the 'accepted' contemporary interpretation[13] and the ICRC 2016 Commentary and the Expert Opinion requested in light of it.[14] This is to enable the analysis of the relationship between CA1 and AWS in its potentially most influential form, that is to say, whether it can require a tether to be included by the supplying state in all AWS it supplies.

As pointed out by the 2016 Commentary, the meaning of the term 'ensure' is to make sure something will occur or in this case will not occur, i.e. violations of the Conventions.[15] Logically, this goes beyond a prohibition on encouraging, aiding or assisting violations of the Convention by parties to a conflict. Therefore, ensuring respect within the meaning of CA1 includes a preventive aspect, whereby the HCPs must take steps to prevent foreseeable violations, both during peace and wartime, which, as mentioned above, is also directed towards other parties such as those in a conflict. The positive obligation also requires that the HCP does 'everything reasonably in their power to […] bring such violations to an end'.[16]

In relation to preventing future violations, there must be a foreseeable risk of them being committed.[17] The actual means by which a state is to carry out this obligation is largely at its discretion, provided the principle of due diligence is adhered to.[18] Hence, the positive external duty to ensure respect is an 'obligation of means', whereby an HCP is not held responsible for a failure of its efforts, provided it did everything reasonably in its power.[19] Consequently, the HCP must first correctly identify foreseeable future violations and then take all measures reasonably in its power to prevent them.

13 Breslin (n 8) 37.
14 Boutruche, Sassoli (n 7) 13.
15 International Review of the Red Cross, *Commentary on the First Geneva Convention: Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field, 2nd edition* [145].
16 Ibid. 154.
17 Ibid. 164.
18 Ibid. 165.
19 Ibid.

The 2016 Commentary goes on to refer to the 'unique position' of influence where an HCP takes part in the arming, training or otherwise equipping of the armed forces of a party to a conflict.[20] If we consider autonomous weapons systems in this context, it is apparent that if an HCP is providing such weapons, it is arguably in a unique position to prevent or end violations as it could reasonably have taken a multitude of steps to increase its influence beforehand, such as placing remote kill-switches on the supplied systems. Arguably, this is the first time the use of physical weapons systems in the physical possession of a state to which it has been supplied can be made conditional on complying with IHL, even if conceivably similar conditions could already in the present be attached to the use of cyber capabilities supplied by another state. Therefore, whereas in the case of conventional human-operated weapons the most the supplying party could do directly is to stop further supply, under the new paradigm the threat could be to make existing systems useless, thus greatly increasing the leverage. This would effectively prevent future violations, at least by those AWS that can be disabled. Which both introduces the importance and leads us to the main topic of this paper: what are the implications of CA1 in relation to an HCP supplying AWS, and is the supplier required to maintain a tether enabling control of those supplied systems?

3. AUTONOMOUS WEAPONS SYSTEMS

Autonomous weapons systems are no longer contained within the realm of science fiction, as already in the present day there are, for example, missile defence systems that can work entirely autonomously. These include the U.S. Aegis command system and the Phalanx Close-in Weapons System (CIWS), which has a mode where it presumes the human operators are incapacitated and it can engage incoming missiles and aircraft on its own.[21] From this example, we may derive the key aspects for defining an autonomous weapons system: a weapons system that is capable of independently identifying and making the decision to engage targets without human intervention, which closely mirrors the U.S. definition of an AWS.[22] There is much discussion regarding the precise definition; however, for the purposes of this paper, we will use the definition whereby a weapons system is autonomous when it can identify, target and engage without human intervention.

The lack of human influence has led to discussions about the 'responsibility gap'[23] regarding AWS, similar to the discussion about liability for self-driving cars and other vehicles. In both cases, the options that are most often discussed are that either the manufacturer or programmers are liable, or the seller, the operator in limited cases,

20 Ibid. 167.
21 Rebecca Crootof, 'Autonomous Weapon Systems and the Limits of Analogy' (2018) HNSJ 9, 59.
22 Ingvild Bode, Hendrik Huess, 'Autonomous Weapons Systems and changing norms in international relations' (2018) Review of International Studies 44, 399.
23 Marcus Schulzke, 'Autonomous Weapons and Distributed Responsibility' (2013) Philosophy & Technology 26, 206.

or the user (such as in the case of neglect that leads to a failure), or even the machine itself.[24] While each has its pros, cons and limitations, the discussion is too complex to attempt to resolve in this paper.

Nevertheless, a few aspects must be discussed in this regard. Firstly, the question of the possibility of human intervention is crucial for the accountability for the actions of the autonomous system. Arguably, if a person has the possibility of influencing the autonomous system, it is not truly autonomous, as that person will be held responsible for failing to prevent the system from malfunctioning. In the case of autonomous vehicles, there are complex legal and ethical questions of whether such a possibility should even be included, as its inclusion would defeat the point of the autonomous vehicle; the human would still have to supervise it, thereby removing the benefit of, for example, sleeping while travelling.[25]

The same will hold true for AWS, but with the added dimension that now the autonomous system can make decisions to specifically end human life. Therefore, in the case of AWS, the pressure to include such safeguards is increased, but this raises further ethical questions; if the AWS is capable of operating unsupervised in a dangerous situation, is it ethical to endanger your own soldiers' lives by placing them inside the system to monitor its operation?

Secondly, it may be an unfortunate reality that not all AWS can be monitored if they are on the offensive, as it may be beneficial from a military point of view that they abstain from unnecessary communications and are as 'radio-silent' as possible, to prevent their location and destruction by the enemy. Hence, it is conceivable that future AWS may not have any human overrides, which would create the 'accountability gap'.[26] This would mean that the supplying state, if it so desires, could distance itself from supplied AWS in a similar way to 'traditional' weapons operated by humans, by stating that the users have the possibility of influencing them.

A further aspect in relation to AWS, which is closely related, is the unprecedented opportunity to include a pre-programmed 'basic moral code', whereby the AWS would simply refuse to comply with certain commands, such as those in clear violation of the Geneva Conventions. This situation is distinct from present reality, where human combatants may harbour hidden 'characteristics' unknown to their commanders, such as hatred of certain ethnicities, a thirst for revenge in the heat of battle, or hidden mental diseases. The possible presence of these hidden characteristics in human combatants is preventable in AWS, where, despite a potential capacity to learn and

24 Alexander Hevelke, Julian Nida-Rumelin, 'Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis' (2015) Science & Engineering Ethics 21, 620-621 & 623-624.

25 Ibid, 619-630.

26 Marcus Schulzke, 'Autonomous Weapons and Distributed Responsibility' (2013) Philosophy & Technology 26, 206.

adapt, the programming of the system could nonetheless include safeguards like Asimov's Laws of Robotics,[27] i.e. absolute prohibitions that underlie all operations.

Due to this possibility, the state supplying and producing AWS has a concrete and unique opportunity to prevent those systems from violating IHL norms, and thus 'ensure respect' for the Geneva Conventions. Potentially, the AWS could even be used as a 'vigilance system', whereby the AWS observing violations of IHL would either store details of those violations in a black box type of storage or send them to the manufacturer or another relevant entity, such as the Protecting Power or even the ICRC. Similarly, the AWS could store all the orders it has received from its human operators in a log, allowing for retroactive tracing of who gave the command and exactly what the command was, thus identifying commands that would have used the AWS to commit violations of IHL. If such features were to be included, non-physical safeguards should be considered, as suggested in the Guiding Principles of a 2019 draft report by the Group of Governmental Experts for the UN Convention on Certain Conventional Weapons (CCW), to prevent, for example, data spoofing that would reduce the utility of such a log and increase uncertainty related to its integrity.[28] All these possibilities hinge on the producer of the AWS including or being required to include such features into their machines, thereby giving further value to defining the obligations of CA1, as these possibilities, if they are technically feasible at the time, could certainly be included in measures reasonably in the power of the HCP supplying the AWS.

Nonetheless, at the time of writing, though many discussions have taken place about legally regulating AWS, especially in the context of the CCW in the form of a pre-emptive ban such as in the case of blinding laser weapons, at present there are no international legally binding instruments on AWS.[29] Therefore, considering that autonomous weapons such as CIWS are already in use, and many research programmes are underway, it is safe to say the legal practice is lagging.[30]

27 Roger Clarke, 'Asimov's Laws of Robotics: implications for information technology' (1993) Computing Milieux, 55.

28 United Nations, 'Draft Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of lethal Autonomous Weapons Systems' <https://www.unog.ch/80256EDD006B8954/(httpAssets)/5497DF9B01E5D9CFC125845E00308E44/$file/CCW_GGE.1_2019_CRP.1_Rev2.pdf> accessed 17 April 2020.

29 Ingvild Bode, Hendrik Huess, 'Autonomous Weapons Systems and changing norms in international relations' (2018) Review of International Studies 44, 398-400.

30 Ibid. 400.

4. INTERACTION OF AWS AND COMMON ARTICLE 1

A. Not all AWS are Created Equal

Autonomous weapons systems are not mentioned in the 2016 Commentary, nor how the obligations of the Article would interact with them. Nonetheless, based on the discussion in the previous section about the nature of AWS, as they can make decisions to engage targets on their own, it is foreseeable that they could do so in violation of IHL norms. Therefore, the positive obligation of preventing violations when there is a foreseeable risk[31] would apply to such AWS systems.

This presumes that the AWS systems in question can cause harm or use lethal force, meaning that a distinction must be made between AWS systems where it is foreseeable that they may cause violations and those that foreseeably could not. It is reasonable to presume that the armed forces will adopt (unarmed) autonomous vehicles such as cars and trucks, but arguably, as these are not designed to have a combat role, they are unlikely to cause violations of IHL in their normal operations. By contrast, the moment an autonomous vehicle is armed, the situation becomes different, as foreseeably the armament could be misused.

The distinction may be even more difficult if we consider the present example of the already autonomous Goalkeeper CIWS system, which can engage missiles and aircraft on its own. First, we must consider that it is a mounted system that is immobile, and so its operation can be closely monitored by humans, even if the people doing the monitoring do not contribute to the decision-making of the system, and the system can be shut down if it malfunctions. Secondly, the system is designed to engage high-speed targets such as missiles and aircraft with the capacity to identify friend or foe (IFF functionality), meaning that it can distinguish between civilian and military aircraft.[32] Thirdly, the system is short-ranged (2000 metres),[33] which in combination with only targeting high-speed objects such as missiles, and its ability to distinguish civilian aircraft, would mean that the foreseeable violations would be limited to engaging a misidentified civilian aircraft that strayed within 2000 metres of the system. Considering the specification of this system, despite it being a lethal AWS as it is capable of destroying aircraft, it is difficult to identify many foreseeable risks in terms of IHL violations, as it is highly unlikely to interact with protected persons under the Geneva Convention and could violate IHL in highly specific scenarios only. By comparison, a mobile airborne autonomous drone engaging in a persistent

31 International Review of the Red Cross, *Commentary on the First Geneva Convention: Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field, 2nd edition*, (2016) [164].

32 Seaforces, 'Goalkeeper close-in weapon systems' <http://www.seaforces.org/wpnsys/SURFACE/Goalkeeper-CIWS.htm> accessed 23 December 2019.

33 Ibid.

campaign of targeted killings[34] would be at a higher risk of foreseeably causing IHL violations, as it could target a variety of ground forces, installations and civilian targets. Consequently, the range of foreseeable violations of IHL that the system is capable of causing is far wider than in the case of an autonomous CIWS system.

Both systems in the above examples can be exposed to cyber threats as they rely on and are operated by computer systems. Hence, it is plausible to consider a scenario where a cyberattack causes the AWS to violate IHL.[35] Although there is currently no obligation on states to foresee and analyse possible misuses of weapons,[36] it may be argued that, given the relative but inherent insecurity of computer systems, it can be reasonably expected that tampering by cyber means will sooner or later take place and affect the normal and expected use of an otherwise legal AWS. While no such binding obligation exists, the topic of cyber security in AWS in the context of non-physical safeguards has been mentioned in the Guiding Principles of a 2019 draft report by the GGE for the CCW Convention as an aspect to consider, thereby suggesting at the very least mounting discussions on the topic that could eventually lead to binding obligations in the future.[37] Potential misuses of AWS by adversaries via exploiting unknown vulnerabilities and resulting in the risk of violations of IHL are hardly foreseeable in advance. However, the same cannot be said about already-known vulnerabilities. Therefore, although analysis of misuse may not be required under IHL or other international law obligations, it is questionable whether the existence of a known vulnerability in an AWS that could potentially lead to violation of IHL would render the risk of that violation foreseeable.

Consequently, the foreseeable risk of violations is highly specific to the type of AWS, and as such, AWS cannot be categorised merely based on their autonomous function or potential lethality, but rather a system-by-system overall risk analysis must be performed. For states party to Additional Protocol I, Article 36 does require that reviews are conducted for each new weapon developed or acquired; however, major military powers such as the United States are not bound by AP I, thereby limiting its reach.[38] Moreover, Article 36 weapons reviews that are conducted are not required to be published and therefore can be subject to secrecy, so arguably this lack of transparency could compromise the effectiveness and truthfulness of the reviews that

34 Michael Carl Haas, Sophie-Charlotte Fischer, 'Evolution of targeted killing practices: autonomous weapons, future conflict and international order' (2017) Contemporary Security Policy 38, 283.

35 Michael N. Schmitt, 'Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics' (2013) Harvard National Security Journal Features, 7.

36 ICRC Commentary on the Additional Protocols, paragraph 1469. Also see Michael N. Schmitt (ed) Tallinn Manual 2.0. on the International Law Applicable to Cyber Operations, 466.

37 United Nations, 'Draft Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of lethal Autonomous Weapons Systems' <https://www.unog.ch/80256EDD006B8954/(httpAssets)/5497DF9B01E5D9CFC125845E00308E44/$file/CCW_GGE.1_2019_CRP.1_Rev2.pdf> accessed 17 April 2020.

38 Natalia Jevglevskaja, 'Weapons Review Obligation under Customary International Law.' (2018) U.S. Naval War College International Law Studies, Vol 94, 209.

are conducted.[39] Nevertheless, a discussion regarding Article 36 of AP I is beyond the scope of this paper and a comprehensive examination thereof would require a separate article to examine.

A state supplying a system like the Goalkeeper CIWS would arguably have to take fewer preventive steps to inhibit the system from causing IHL violations than a state supplying an autonomous 'killer-drone'. The actual content of the obligations under CA1 would be different based on the types of AWS supplied, and could not be mapped precisely in the abstract. However, it is possible to state abstractly that the HCP should take all measures in ensuring that the AWS cannot cause the foreseeable violations of IHL specific to that system. Such measures should include a misuse risk assessment by identifying and appropriately addressing at least the known cyber vulnerabilities that can lead to violations of IHL.

B. The External Positive Obligation of Common Article 1

Under CA1, HCPs have the positive obligations of both preventing future violations and stopping ongoing violations by a party to a conflict. Consequently, AWS provide the unprecedented opportunity to definitively pre-programme a set of rules that the physical weapons system must follow, such as to prevent violations of IHL. Of course, considering the complexity of both practical situations in a conflict and the legal framework, the correct course of action can be difficult to determine and there has been doubt expressed about whether AWS can ever operate within the correct manner from an IHL point of view.[40] However, arguably that is dependent on the type of system, as outlined above in 4.A.

It would be a gross oversimplification to reduce the situation to programming the system with a simple set of rules such as 'never target non-military infrastructure' or 'never cause the death of a civilian' to definitively prevent violations. While both are in theory protected, in practice the situation may be more complicated and would not necessarily involve a violation of IHL, depending on the proportionality and the military advantage gained. For example, a bridge can be entirely a civilian structure, however, the military advantage of destroying that bridge may justify its destruction, thus abstractly transforming it from a civilian structure to a military target.[41] Similarly, in the case of a targeted killing campaign, if a high-ranking enemy is found who is in the presence of a civilian and a decision to engage would end both their lives,

39 International Review of the Red Cross, *Commentary on the Protocol Additional to the Geneva Conventions of 12 August 1949 and relating to the Protection of International Armed Conflicts (Protocol 1)*, (1987), 1470.

40 Max van Kralingen, 'Use of Weapons: Should We Ban the Development of Autonomous Weapons Systems?' (2016), The International Journal of Intelligence, Security and Public Affairs, 18:2, 137.

41 ICRC, 'Practice Relating to Rule 10. Civilian Objects' Loss of Protection from Attack' <https://ihl-databases.icrc.org/customary-ihl/eng/docs/v2_rul_rule10> accessed 23 December 2019.

conceivably considerations of military advantage and proportionality could justify the killing of the civilian alongside the high-ranking commander.[42]

Both cases highlight that commands that appear almost like a tautology such as 'never kill or cause the death of a civilian' are not always realistically possible to include as overruling laws, in a manner similar to Asimov's Laws of Robotics. Consequently, the task of pre-programming an AWS to such an extent that under absolutely no circumstances could it violate IHL is a herculean task. The supplier of an AWS could likely never eliminate the chance of their AWS causing violations purely based on its programming. Of course, if such a technological feat is possible, feasibly CA1 would require that the supplied AWS would be included with such programming as it would be a measure reasonably in the power of the supplying state. However, we must be realistic and assume it is not possible, at least for all systems, for the near future.

Therefore, from the above conclusion we arrive at the second possibility that could potentially be required under CA1, the question of whether or not the supplying HCP has an obligation to retain the possibility of influencing the AWS or monitoring its activity.

C. To Tether or Not to Tether?

The possibility of influencing the actions and behaviour of AWS using remote-control raises the possibility of HCPs meeting the positive obligation of CA1 by taking control of their supplied AWS. This question is similar to that which has been posed regarding encryption: whether backdoors should be provided to give authorities access.[43] In the case of AWS, the discussion will have the added life-and-death dimension whereby if a backdoor is included and the system is hacked, lives could be lost. The presence of a backdoor also increases the number of actors potentially able to commit IHL violations with the AWS, should a third party be able to hijack the system by exploiting the backdoor. To a degree, this risk could be reduced by limiting the backdoors to only disabling the AWS, which if breached would at least not cause violations, but would hamper the functionality of the AWS considerably.

There could be no better or more immediate way of preventing violations by AWS used by a party to a conflict than remotely disabling those systems being misused. Therefore, from a compliance perspective, the ability to remotely monitor and influence would ensure the respect for the Geneva Convention and other applicable IHL, even if it is a double-edged sword due to the risk of unauthorised access. Several other considerations should be considered when determining whether tethering the AWS should be required as a means of fulfilling the obligations under CA1.

42 ICRC, 'Practice Relating to Rule 14. Proportionality in Attack' <https://ihl-databases.icrc.org/customary-ihl/eng/docs/v2_rul_rule14> accessed 23 December 2019.

43 Ronald Rivest, 'Case against regulating encryption technology' (1998) Scientific American, 116-117.

First, let us consider the untethered model, whereby, to begin with, the supplying state severs or does not include all possibility of influencing the supplied system once it has been supplied to another state. The supplying state would be entirely unable to monitor or direct its activities in the future. This would render AWS akin to traditional human-operated weapons systems whose country of origin has no control over how they are used after handing them over. Consequently, the supplying state would have to resort to the traditional means of influence such as diplomatic pressure, economic sanctions and refusing to supply the party in the future.[44]

Under this untethered model, the introduction of AWS changes less in how the HCPs comply with the obligation to 'ensure respect' under CA1. The only meaningful improvement would be the programming of the AWS aimed at preventing the misuse of the AWS, which would be included under the measures that HCPs can reasonably take to prevent foreseeable violations. This would likely not cover all possible situations where violations can occur, and hence would likely not be a panacea. Nonetheless, when compared to the present where the compliance or non-compliance of non-autonomous weapon systems is entirely at the mercy of their crews,[45] it would be an improvement.

The second possibility is the 'tethered' model, which could be described by analogy as a 'Swiss mercenary of old' model. If the supplying state maintains some form of connection, be it the capacity to monitor the activity, direct the activity or have a remote 'kill-switch' for the AWS, the AWS is not truly an asset of the state it has been supplied to, but rather something of a cyber mercenary'. In this sense it is similar to the 'Swiss mercenaries of old', whose service came with conditions in regard to their state of origin (Switzerland) such as that they might be recalled if the Swiss confederacy were to come under attack.[46] Consequently, a prudent user of the Swiss mercenaries would have understood that they could not be entirely relied on in all circumstances. Similarly, if the AWS were tethered to its state of origin it could not necessarily be relied on in all circumstances, such as when those AWS were used to cause violations of IHL or conflict with the supplying state. Especially if there were a conflict with the state supplying the AWS, the user might find that those systems had 'turned traitor', adding a whole new level to cyber warfare, and as such putting them at a great military disadvantage.

That is to say, the AWS could never be 'fully trusted' in the same way as the 'Swiss mercenaries of old', who, while entirely under the command of the local armed forces,

44 Knut Dormann, Jose Serralvo, 'Common Article 1 to the Geneva Convention and the obligation to prevent international humanitarian law violations' (2014) ICRC 96, 725-726; International Review of the Red Cross, *Commentary on the First Geneva Convention: Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field, 2nd edition*, (2016) 181.

45 Hin-Yan Liu, 'Categorization and legality of autonomous and remote weapons systems' (2012) ICRC 886, 629-630.

46 John McCormack, *One Million Mercenaries: Swiss Soldiers in the Armies of the World* (Pen and Sword 1993), 62.

would nonetheless have a link to their state of origin. Similarly, tethered AWS would have a remote cyber link to the supplying state, which may be activated at any time, thus transforming them into a 'cyber mercenary' from the supplying state. Therefore, the armed forces of a state buying AWS might be compromised by the presence of these 'cyber mercenaries' in their ranks, which would enable the supplier of those AWS to retain both political and military leverage over the state using those systems. Naturally, this analogy is restricted by the use of the term 'mercenary', as there is a risk of confusion with the term's present legal meaning under which a 'mercenary' does not retain any link to their state of origin.[47] Consequently, this 'cyber mercenary' would arguably require a new term without pre-existing definitions or prejudices. In this vein, a portmanteau between 'autonomous' and 'mercenary' could be used, such as 'autocenary', which could be defined as 'an autonomous weapons system that is tethered to its state of origin or production by means that enable monitoring or remote control'. Nevertheless, despite its limitations, the term 'cyber mercenary' will be used for the purposes of this paper.

Nevertheless, the tethering of AWS to the supplying state would solve one of the key questions of supplying weapons: what if they are ever used against the supplier? For on the one hand, the supplier wants to supply inferior systems so that they cannot compete with their own, but at the same time they must be better than the competing systems which would otherwise be chosen. Maintaining control would give the best of both worlds to the supplier: the systems can be as effective as possible, as the supplier knows that if ever it was used against them, they could disable or control it. Merely being able to monitor its use would allow the supplier to spy on the supplied state's armed forces, and as such gain valuable intelligence. If we accept that only major military powers will be able to produce and develop their own AWS, tethering them to the supplier would multiply their leverage over states that are forced to purchase foreign systems and are thus left with an unreliable military full of 'cyber mercenaries'. The leverage gained by such a tether would be both military and political, as not only does the supplying state have a measure of control over the military of the supplied state, but also political capital. This control could be used to ensure favourable relations with the supplying state by exploiting that leverage given by the tethered AWS.

However, it must equally be remembered that if the supplier can remotely access the AWS, conceivably so could a third party; thus the presence of tethering will increase the vulnerability of the systems to cyber-attack by third parties. This threat is especially elevated by the fact that if such a tether is required by law, third party actors will know that it must be present, therefore justifying a significant investment into attempting to exploit such a tether and the leverage over the military of the supplied state brought with it. If no tether is required, third party actors would have to consider

47 International Convention against the Recruitment, Use, Financing and Training of Mercenaries, 4 December 1989, Article 1 (1) (e) and 1 (2) (d).

if such a tether even exists, and thereby the incentive to invest significant resources into exploiting a potential tether would be reduced.

While tethering the systems to the supplying state might appear the most tempting option to fulfil the positive external obligation under CA1, if such a tethering were to be required it would have significant undesirable consequences for any state purchasing such systems. Therefore, it would be prudent not to be naïve when the tethered model is being advocated under the guise of added or assured compliance with the obligations of both IHL, and especially, CA1. Nonetheless, it is equally possible that hidden backdoors and overrides can never be conclusively eliminated, regardless of whether or not this would be required by CA1, as the potential leverage is so tempting.

5. CONCLUSION

The relationship of the positive external obligation of CA1 and AWS can take on a variety of directions; however, the key factor of the relationship is the question of tethering the supplied AWS so that the supplying state can 'ensure respect', as required by CA1, in all circumstances. Certainly, from a legal point of view, a compelling case can be made for requiring such tethering based on the need for HCPs to do 'everything reasonably in their power to prevent and bring such [IHL] violations to an end'[48] under the positive obligation of CA1. Consequently, provided that such a tether is technically feasible, it would be within the reasonable power of the supplying state to include such a backdoor for access, and would significantly aid in preventing both future and ongoing violations.

The choice, however, is more difficult and complex, as the trade-off is either potentially sacrificing compliance by not requiring the tethering, or potentially compromising the armed forces of the supplied states with these autonomous 'cyber mercenaries'(autocenaries) in their ranks in exchange for added compliance. The presence of tethering would also likely significantly increase the risk of the AWS being hijacked by a third party, thereby further adding to the cyber security concerns of the systems. Requiring the tethering of the AWS would have significant political and military implications by further increasing the power of the states supplying AWS, and the potential military leverage gained by cyber warfare for third parties seeking to exploit the tether would be increased.

Moreover, it must be kept in mind that not all AWS are the same and involve similar foreseeable risks of committing violations of IHL. Therefore, the question of 'to tether

48 International Review of the Red Cross, *Commentary on the First Geneva Convention: Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field, 2nd edition*, 154.

or not to tether' could be broken down to a case-by-case basis, wherein, for example, an AWS that has a relatively low risk of causing violations, such as a stationary missile defence system, would not be under a tethering requirement, but a higher-risk 'killer-drone' would be. Under such a system-by-system model, however, the legitimate concern can be raised that, if one state supplies both tethered and untethered AWS, how is the receiving state ever going to silence the doubt that on the 'untethered' systems, the tethers are merely hidden? Consequently, further discussions and contemplations are required on the matter, for conceivably at present the positive obligation of CA1 could be used to justify such a tethered system, as it would ensure a higher degree of compliance and respect for the Geneva Conventions and other applicable IHL.

The positive external obligation of CA1 has implications for the use and development of AWS and the states supplying them. The identified primary key issue arising from the relationship between CA1 and AWS is the question of the tethering of AWS to the state of origin. However, as AWS can take a variety of forms with different risk profiles, it is difficult to provide an all-encompassing answer to whether tethering would be appropriate in every case. This uncertainty is compounded by the additional political and military ramifications of tethering, as it would likely result in an increased power imbalance between the state using the AWS and the supplying state. Therefore, in conclusion, the positive external obligation of CA1 has serious implications for AWS in potentially requiring tethering to the supplying state, a question which is best approached on a system-by-system basis owing to the diversity of AWS and their differing risk profiles.

REFERENCES

Primary Sources

1. Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field (First Geneva Convention), 12 August 1949, 75 UNTS 31.

Secondary Sources

2. International Review of the Red Cross, *Commentary on the First Geneva Convention: Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field, 2nd edition*, (2016).
3. International Review of the Red Cross, *Commentary on the Protocol Additional to the Geneva Conventions of 12 August 1949 and relating to the Protection of International Armed Conflicts (Protocol 1)*, (1987).
4. International Review of the Red Cross, *Commentary on the Additional Protocols*, (1987).
5. ICRC, 'Practice Relating to Rule 10. Civilian Objects' Loss of Protection from Attack' <https://ihl-databases.icrc.org/customary-ihl/eng/docs/v2_rul_rule10> accessed 23 December 2019.
6. ICRC, 'Practice Relating to Rule 14. Proportionality in Attack' <https://ihl-databases.icrc.org/customary-ihl/eng/docs/v2_rul_rule14> accessed 23 December 2019.
7. Seaforces, 'Goalkeeper close-in weapon systems' <http://www.seaforces.org/wpnsys/SURFACE/Goalkeeper-CIWS.htm> accessed 23 December 2019.
8. Bode, I., Huess, H. 'Autonomous Weapons Systems and changing norms in international relations' (2018) Review of International Studies 44, 393-413.

9. Boutruche, T., Sassoli M, 'Expert Opinion on Third States' Obligation vis-à-vis IHL Violations under International Law, with a special focus on Common Article 1 to the 1949 Geneva Conventions' < https://www.nrc.no/resources/legal-opinions/third-states-obligations-vis-a-vis-ihl-violations-under-international-law/> accessed 21 February 2020.
10. Breslin, A. 'Reflections on the Legal Obligation to Ensure Respect' (2017), Journal of Conflict and Security Law 22(1), 5-37.
11. Cameron, L., Demeyere, B., Henckaerts, J-M., La Haye, E., Niebergall-Lackner, H.'The updated Commentary on the First Geneva Convention – a new tool for generating respect for international humanitarian law' (2015), ICRC 97,1209-1226.
12. Clarke, R. 'Asimov's Laws of Robotics: implications for information technology' (1993) Computing Mileux, 53-61.
13. Crootof, R.'Autonomous Weapon Systems and the Limits of Analogy' (2018) HNSJ 9, 51-83.
14. Dormann, K., Serralvo, J.'Common Article 1 to the Geneva Convention and the obligation to prevent international humanitarian law violations' (2014) ICRC 96, 707-736.
15. Haas, M., Fischer, S-C. 'Evolution of targeted killing practices: autonomous weapons, future conflict and international order' (2017) Contemporary Security Policy 38, 281-306.
16. Hevelke, A., Nida-Rumelin, J. 'Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis' (2015) Science & Engineering Ethics 21, 619-630.
17. Jevglevskaja, N. 'Weapons Review Obligation under Customary International Law.' (2018) U.S. Naval War College International Law Studies, Vol 94, 186-221.
18. Liu, H. 'Categorization and legality of autonomous and remote weapons systems' (2012) ICRC 94, 627-652.
19. United Nations, 'Draft Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of lethal Autonomous Weapons Systems' <https://www.unog.ch/80256EDD006B8954/(httpAssets)/5497DF9B01E5D9CFC125845E00308E44/$file/CCW_GGE.1_2019_CRP.1_Rev2.pdf> accessed 17 April 2020.
20. van Kralingen, M. 'Use of Weapons: Should We Ban the Development of Autonomous Weapons Systems?' (2016), The International Journal of Intelligence, Security and Public Affairs, 18:2, 132-156.
21. McCormack, J., *One Million Mercenaries: Swiss Soldiers in the Armies of the World* (Pen and Sword 1993).
22. Rivest, R. 'Case against regulating encryption technology' (1998) Scientific American, 116-117.
23. Schmitt, M. 'Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics' (2013) Harvard National Security Journal Features, 1-37.
24. Schulzke, M. 'Autonomous Weapons and Distributed Responsibility' (2013) Philosophy & Technology 26, 203-219.
25. Zych, T. 'The Scope of the Obligation to Respect and to Ensure Respect for International Humanitarian Law', (2009) Windsor Yearbook of Access to Justice 27, 251-270.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Cyber Weapons Review in Situations Below the Threshold of Armed Conflict

Ivana Kudláčková*
Researcher
Faculty of Law
Masaryk University
Brno, Czech Republic
ivana.kudlackova@law.muni.cz

David Wallace
Professor
Department of Law
United States Military Academy
West Point, New York, US
david.wallace@westpoint.edu

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

Abstract: The use of cyber weapons raises many issues, one of which is the scope of legal requirements affecting the legal review of cyber weapons under Additional Protocol I and customary international law. This paper explores the review of cyber weapons intended for use below the threshold of armed conflict

As the line between war and peace is often increasingly blurred and the majority of cyber incidents are below the threshold of armed conflict, the laws and principles of international humanitarian law do not apply. In this paper, we engage in a scenario-based thought experiment exploring the legal framework affecting the use of cyber weapons outside armed conflict. In such situations, the well-known article 36 of Additional Protocol I and customary international law are not triggered. As a result, there is no explicit legal obligation to conduct a cyber weapons review in situations when cyber weapons are deployed in situations falling below the threshold of armed

* Ivana Kudláčková's and Jakub Harašta's contributions to this paper were supported by ERDF 'CyberSecurity, CyberCrime and Critical Information Infrastructures Center of Excellence' (No. CZ.02.1.01/0.0/0.0/16_019/0000822).

conflict. Our starting point is that even though international humanitarian law is not applicable, the use of cyber weapons is not completely unregulated.

In the paper, we search for answer to following research question: what are the legal requirements for weapons review in situations where their intended use is for situations below the threshold of armed conflict? We identify the black-letter legal framework and explore the state practice of NATO member states where available.

The paper argues that there are many obligations to be considered when deploying cyber weapons in situations below the threshold of armed conflict. The conclusion is that there is no obligation to conduct a review outside Article 36 of Additional Protocol I. That being said, there are definitely policy benefits in conducting broader software assessment to ensure respect to international law obligations of a state.

Keywords: *cyber weapons, software, legal review, art. 36 of Additional Protocol I, human rights*

1. INTRODUCTION

The regulation of cyber weapons under international law has been an unsettled issue, not only among international lawyers but also among information technology specialists and political and security researchers. This uncertainty presents a challenge to reconsider the existing norms of international law, especially the obligation to conduct a weapons review imposed by Article 36 of Additional Protocol I (API) and customary international law. Legal review of cyber weapons has been already discussed in, amongst other places, the *Tallinn Manual 2.0* and the Cyber Law Toolkit. Understandably, both focused on the weapons review requirement under international humanitarian law (IHL). In our opinion, however, this approach does not fully reflect the cyber reality. In this paper, we explore the issue of weapons review beyond Article 36 of API and examine other possible legal regimes to account for legal requirements that appear elsewhere in the conflict classification framework.[1]

This paper seeks to answer the following research question:

What legal requirements need to be considered when deploying cyber weapon in situations below the threshold of armed conflict?

1 Compare David A. Wallace and Christopher W. Jacobs, 'Conflict Classification and Cyber Operations: Gaps, Ambiguities and Fault Lines', (2019) 40 *U. Pa. J. Int'l L*, 643.

The paper is structured as follows. First, we encapsulate existing definitional approaches to cyber weapons and introduce our working definition. Second, we present a hypothetical scenario with escalating conflict between two fictional States – scenario contains both cyber and non-cyber events that drive escalation towards armed conflict. Third, these incidents are explored through the lens of various legal regimes, such as derogation of human rights, issues of sovereignty and non-intervention, and the use of force, armed attack and armed conflict. Finally, we discuss the existing connection between limits imposed on the use of cyber weapons by international public law in general, hence reaching beyond the narrow scope of Article 36 of API.

2. CYBER WEAPONS: WORKING DEFINITION AND WEAPONS REVIEW

A. Cyber Weapons

Given the various technical, legal, security and policy aspects of the term *cyber weapons*, it is highly unlikely that a universally accepted definition will ever be reached. That being said, reaching at least a working definition makes the issue more accessible for discussion. The term *weapon* carries normative meaning pointing us directly to Article 36 of API. Automatically, it triggers the requirement to conduct a formalised weapon review. Therefore, we use the term software for scenarios below the threshold of armed conflict and we reserve the term cyber weapon only for the context of international armed conflict. Our decision directly stems from the wording used in Rule 103 of the *Tallinn Manual 2.0* and from some of the works mentioned below.

Generally, scholars trying to define cyber weapons follow two trends. The first group focuses on the intended target of the cyber weapon and on its ability to cause damage.[2] Damage is crucial here, as some authors acknowledge that without the ability to cause damage, even highly invasive techniques such as data exfiltration do not constitute a cyber weapon.[3] We are proponents of the concept that data is an object and might be qualified as a military objective.[4] However, we recognise that this is a very controversial and unsettled issue. The second group simply refers to cyber incidents without really intending to provide a clear definition of the term cyber weapon. Some authors mention Stuxnet, the DDoS attacks on Estonia in 2007 or the use of

2 Peeter, Lorents and Rain Ottis, 'Knowledge Based Framework for Cyber Weapons and Conflict', (2010) *Conference on Cyber Conflict Proceedings* 129, 139. Amit. K. Maitra, 'Offensive cyber-weapons: technical, legal, and strategic aspects', (2015) 35 *Environment Systems and Decisions* 169, 179. Thomas Rid and Peter McBurney, 'Cyber-Weapons', (2012) 157 *The RUSI Journal* 6, 7.

3 Jacqueline Eggenschwiller and Jantje Silomon, 'Challenges and opportunities in cyber weapon norm construction', (2018) 12 *Computer Fraud & Security* 11, 12. Sami Zhioua, 'The Middle East under Malware Attack Dissecting Cyber Weapons', (2013) *IEEE 33rd International Conference on Distributed Computing Systems Workshops Proceedings* 11, 11.

4 Compare Kubo Mačák. 'Military Objectives 2.0: The Case for Interpreting Computer Data as Objects under International Humanitarian Law', (2015) 48 *Israel Law Review* 55.

the malware Shamoon against Saudi Aramco in the same breath.[5] If those incidents are not followed by in-depth analysis with an aspiration towards understanding the term cyber weapon and its normative consequences, it presents a threat of undesirable simplification that floods the issue of cyber security.

The International Group of Experts (IGE) drafting the *Tallinn Manual 2.0* dedicated Rule 103 not only to weapons, but also more broadly to means and methods of cyber warfare in general. Cyber weapons are understood to be 'cyber means of warfare that are used, designed, or intended to be used to cause injury to, or death of persons or damage to, or destruction of 'objects'.[6] Furthermore, the IGE distinguished between cyber weapons and cyber systems. A weapon is one of the aspects of a cyber system and is used to 'cause damage or destruction to objects or injury or death to persons'.[7] Given the scope and aim of the *Tallinn Manual 2.0*, these definitions stem mainly from IHL and reflect predominantly Article 36 of API. The definition of cyber weapons is thus closely tied to that of the cyber attack in Rule 92 of the *Tallinn Manual 2.0*. In this view, cyber weapons are intended to execute cyber attacks.

However, as the nature of interstate interaction and possible conflicts evolve, we believe broader considerations are in place. Cyber systems can be used to deliver harmful software to targeted systems. Different payloads can lead to different harmful consequences. However, these consequences may not be so dire as to justify the use of the term 'weapon'; indeed, we believe the current over-use of the term cyber weapon is harmful and obfuscates the discussion. Hence, we take into consideration the cyber systems used to deliver harmful software. Some of the harmful software may ultimately be labelled a cyber weapon. We believe this allows for a more nuanced discussion regarding the existing legal requirements and respects that weapon is just an aspect of a cyber system.[8] As is evident from Figure 1, cyber systems can be used to deliver software into particular targeted devices (different payloads) and only some payloads can be considered cyber weapons. Cyber systems are made up of general infrastructure (operators, means of payload delivery, command and control servers) and additional payloads serving specific purposes. The effects of these payloads may or may not have physical consequences. Some of these payloads may be considered cyber weapons under existing law. However, elucidation of the exact nature of those consequences is not the purpose of this paper. In Figure 1, we do not aspire to provide a universal scheme, but rather to suggest that some sort of a review needs to be conducted, not only in case of use of cyber weapons, but also in case of use of harmful software.

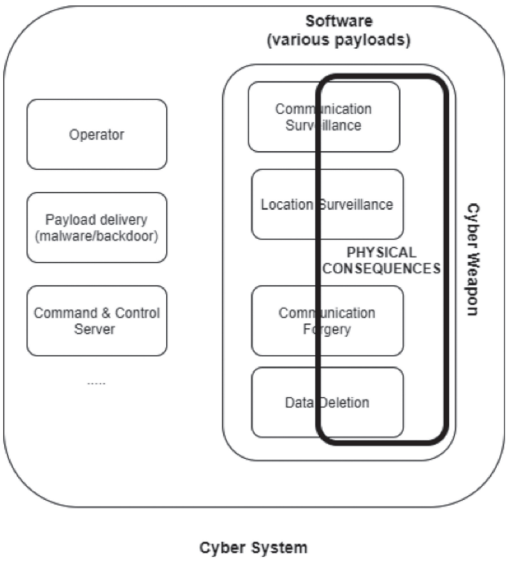
5 Ivanka Barzaszka, 'Are cyber-weapons effective?' (2013) 158 *The RUSI Journal* 48, 48. Gregory D. Koblentz and Brian M. Mazanec, 'Viral Warfare: The Security Implications of Cyber and Biological Weapons', (2013) 32 Comparative Strategy 418, 423. Jeffrey Carr, 'The misunderstood acronym: Why cyber weapons aren't WMD', (2013) 69 *Bulletin of the Atomic Scientists* 32, 34.

6 *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (CUP 2017) ('*Tallinn Manual 2.*0') 452.

7 Ibid.

8 Ibid.

FIGURE 1. REPRESENTATION OF THE RELATIONSHIP BETWEEN A CYBER SYSTEM, A SOFTWARE AND A CYBER WEAPON.



B. Weapons Review

Prohibitions and limitations on weapons are woven deeply into the fabric of IHL[9] and the principles and rules of IHL that regulate weapons are layered.[10] At the broadest level, some general principles and rules apply to all weapons under IHL.[11] Some weapons cannot be directed at a military objective or combatants and would be prohibited because they are inherently indiscriminate. The German V1 rockets used in World War II and the Scud missiles launched by Iraq during the First Gulf War of 1990-91 are examples of such weapons.[12] Beyond the general rules and principles, some treaties regulate or ban specific weapons or classes of weapons such as cluster munitions, landmines, chemical and biological weapons, incendiary weapons and blinding lasers. Finally, Article 36 of API requires State parties to do as follows:

> In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.[13]

9 Gary D. Solis, *The law of armed conflict: international humanitarian law in war* (2nd edn, CUP 2017) 5.
10 Robert Kolb and Richard Hyde, *An introduction to the international law of armed conflicts* (Hart Publishing, 2008) 153.
11 Ibid.
12 UK Ministry of Defence, *The Manual of the Law of Armed Conflict* (Ministry of Defence, United Kingdom, 2004) 104.
13 Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflict (Protocol I), 8 June 1977.

This Article is also reflected in the *Tallinn Manual 2.0* as Rule 110 lit. (b). The IGE was divided on the question of whether Article 36 reflected customary international law or whether it is only applicable to States that have ratified API.[14] Moreover, Rule 110 is not completely exhausted by Article 36 of API, but also contains lit. (a). This rule sets out a customary obligation to ensure that applies to all States and requires them to ensure that the cyber means of warfare that they acquire or use comply with the rule of the law of armed conflict. In our opinion, some issues arise.

First, the nature of a legal review in lit. (a) is unsettled.[15] It remains questionable whether mere advice of a legal advisor on deployment and use satisfies this requirement. The IGE considered it sufficient,[16] taking a practical perspective, as the legal advisor might be the only available option.[17] Regarding lit. (b), there is an obligation to conduct a formal legal review[18] but it is not specified how the review mechanism should be established.[19] Countries such as the United States, the United Kingdom, Belgium, the Netherlands, Norway, Sweden, Australia, France or Germany already have established procedures of legal review for new weapons,[20] but there is no duty to disclose these mechanisms.[21]

Second, the issue of whether a State is party to an armed conflict is not the decisive factor for legal review.[22] Thus, States should carry out a legal review in advance. In this paper, we discuss whether we could imply the same for situations that are below the threshold of armed conflict. The deployment of specific software might trigger armed conflict, and the legal classification of conflict might only be specified after a lapse of time, based on facts of the conflict and further investigation. We therefore believe that software review in broader terms reflects the *ratio* of the existing legal framework.

3. BACKGROUND FOR SCENARIOS

For the purpose of further discussion, we present the following scenario involving the hypothetical escalation of conflict between two fictional States. Berylia and Crimsonia

14 *Tallinn Manual 2.0*, supra n. 6, 465. Compare Natalia Jevglevskaja, 'Weapons Review Obligation under Customary International Law', (2018) 94 *INT'L L. STUD* 186.
15 International Cyber Law: Cyber Law Toolkit. Scenario 10: 'Cyber Weapons Review', <https://cyberlaw.ccdcoe.org/wiki/Scenario_10:_Cyber_weapons_review> [accessed 17 December 2019].
16 *Tallinn Manual 2.0*, supra n. 6, 465.
17 William H. Boothby, *Weapons and the Law of Armed Conflict* (2nd edn, OUP 2016), 341.
18 Compare Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977, Commentary 1987, par. 1970.
19 A Guide to the Legal Review of New Weapons, Means and Methods of Warfare, 20, <https://e-brief.icrc.org/wp-content/uploads/2016/09/12-A-Guide-to-the-Legal-Review-of-New-Weapons.pdf> [accessed 17 December 2019].
20 William H. Boothby, supra n. 17, 343.
21 Supra n. 18.
22 Supra n. 15.

are neighbouring countries.[23] For the purpose of an applicable international legal framework, Berylia is a signatory to the European Convention on Human Rights (ECHR) and API.

One of the Berylian regions directly neighbouring Crimsonia is historically disputed. Citizens of Berylia living in this region align themselves with the nationality that is dominant in Crimsonia. These citizens organise themselves into a political organisation, Crimson Home. The ultimate goal of Crimson Home is cessation from Berylia and incorporation into Crimsonia. Crimson Home intends to reach this goal through a political referendum.

For various reasons, including heightened geopolitical and regional ambitions, both States are prone to escalation of conflict through various triggering events of a cyber and non-cyber nature. These will be described below.

Before the conflict, Berylia had developed cyber capabilities to be able to collect, disrupt and potentially destroy data which adversaries rely upon. For this purpose, the Berylian government procured and developed cyber capabilities allowing the delivery of a harmful payload to target devices. Operators from Berylian law enforcement and armed forces are able to target specific networks or a specific range of IP addresses. Malware can be used to infect targeted devices and obtain sufficient rights to allow the remote delivery of a harmful payload to different components of an operating system. This payload includes modules allowing surveillance of communication, tracking of movement, issuance of counterfeit messages or erasure of data stored on the device. We will refer to this cyber system as Berylian Malware (BERM).

4. STATE VS. CITIZENS

1) Scenario

The first part of our scenario observes a deteriorating relationship between Berylia and Crimsonia. Crimson Home, actively seeking to secede from Berylia, is heavily financed from Crimsonia. The Crimsonian government, despite numerous allegations, has never admitted to supporting Crimson Home. However, finances pouring into Crimson Home originate, according to Berylian intelligence, from Crimsonian companies identified as shell companies used by the Crimsonian government.

After the Berylian government refuses to hold a referendum in conjunction with national elections, Crimson Home heightens its activity. Targeted ads sponsored by Crimson Home aim to incite tension between citizens living in the disputed region and the central Berylian government. This eventually leads to a series of rallies and

23 We follow the naming convention of fictional States used in Locked Shields exercises. However, this in no way implies any endorsement of our paper from any of the Locked Shields organisers.

protests. Social unrest results in small-scale riots, localised violence and spontaneous attacks on election officials and polls. No fatalities are reported, and a number of injured participants is limited to a minimum. Law enforcement agencies use tear gas to disperse the most stubborn protesters. This is directly followed by a series of arrests of people either directly participating in riots or suspected of organising and inciting them. Crimson Home is targeted by BERM. Payload effects include the surveillance of communications and tracking the movement of high-profile members of the organization.

The situation escalates when Crimson Home members organise bombings in the disputed region. These attacks are aimed mainly at buildings representing the central government, the legislature and the courts. The death toll quickly rises into the hundreds. This surge of violence is unprecedented and surprising to the Berylian government. Berylia responds by requiring Crimsonia to cease financing Crimson Home. At the same time, the Berylian government launches large scale operations involving law enforcement agencies as well as a limited deployment of Berylian armed forces in the disputed region. Tension in the region continues to rise. Crimson Home is further targeted by BERM. Payloads still conduct surveillance of communication and tracking of movement. After the bombings, the scale of BERM deployment is increased, and all known members of Crimson Home are targeted.

2) Legal Qualification

Understandably, any State must be aware of its human rights obligations stemming from international treaties. Berylia is obliged to secure rights and freedoms stemming from the ECHR. Nonetheless, Article 15 of ECHR[24] allows derogation from such an obligation. To achieve this, Berylia needs to determine whether a series of rallies, protests and riots is 'an exceptional situation of crisis or emergency, which affects the whole population and constitutes a threat to the organised life of the community of which the State is composed'.[25]

The fact that these events take place only in the disputed region is not an obstacle because 'a crisis which concerns only a particular region of the State can amount

24 Article 15 - Derogation in time of emergency
'1. In time of war or other public emergency threatening the life of the nation any High Contracting Party may take measures derogating from its obligations under [the] Convention to the extent strictly required by the exigencies of the situation, provided that such measures are not inconsistent with its other obligations under international law.
2. No derogation from Article 2, except in respect of deaths resulting from lawful acts of war, or from Articles 3, 4 (§ 1) and 7 shall be made under this provision.
3. Any High Contracting Party availing itself of this right of derogation shall keep the Secretary General of the Council of Europe fully informed of the measures which it has taken and the reasons therefore. It shall also inform the Secretary General of the Council of Europe when such measures have ceased to operate and the provisions of the Convention are again being fully executed'.

25 European Court of Human Rights, *Lawless v. Ireland (No. 3)*, application no. 332/57, 1 July 1961, para. 28.

to a public emergency threatening ”the life of the nation”’.[26] The determination of the situation as a state of emergency is left to the State as a matter of margin of appreciation.[27] A State is not allowed to go beyond what is strictly required by the exigencies of the situation. Whether surveillance of communication and tracking of movement of high-profile members of an organisation complies with this requirement may be assessed against a set of factors based on judicial decisions of the European Court of Human Rights.[28] Assessment of the deployment of BERM when the situation escalates might be clearer if Berylia determines such acts as terrorism, as terrorism meets the standard of a public emergency.[29]

There are also other requirements, but their in-depth analysis is not relevant to this scenario. Therefore, prior to deployment of BERM, Berylia should ensure that its actual deployment will not violate the human rights of its citizens. This could be done by conducting a legal review of BERM against relevant legal obligations and possible derogations in a state of emergency. It is worth noting that clauses similar to Article 15 of ECHR exist within the International Covenant on Civil and Political Rights (Article 4) and the American Charter on Human Rights (Article 17).

5. STATE VS. STATE

A. Sovereignty

1) Scenario

Our scenario follows further escalation. Despite Berylia’s freezing bank accounts of the Crimson Home and its high-profile members as a part of ongoing counter-terrorism operation, Berylian intelligence confirms that Crimsonia has not stopped financing Crimson Home. Financing is now provided by couriers crossing the border from Crimsonia with large sums of cash. Berylian intelligence reports a strong suspicion that weapons are also being transported to Berylia as the Crimsonian government strengthens its support for Crimson Home. BERM is deployed to target any device that connects to specific cell towers located near the border. Payload activities include the surveillance of communication and of movement. However, selected individuals are targeted by harmful payloads allowing suppression of outgoing communication from their devices.

26 European Court of Human Rights, *Ireland v. the United Kingdom*, application no. 5310/71, 18 January 1978, para. 205.

27 European Court of Human Rights. Guide on Article 15 of the Convention. Derogation in time of emergency, para. 11.
< https://www.echr.coe.int/Documents/Guide_Art_15_ENG.pdf> [accessed 17 December 2019].

28 Supra n. 27, para. 21.

29 Supra n. 27, para. 12.

2) Legal qualification

Interstate relations come into consideration at this point. With couriers crossing borders with cash and potentially with weapons, Berylia might decide to consider this situation as a violation of its sovereignty. It seems appropriate to refer to the well-known *Island of Palmas* arbitral award where a definition of sovereignty was proposed,[30] the basic components of which were further developed in Article 2(4) of the UN Charter with key components of territorial integrity and political independence. Berylia might also consider whether to perceive sovereignty as a rule or as a principle. This is still a matter of debate, not only in academia,[31] but also in state practice.[32] It is important to note that Berylia needs to attribute the conduct of couriers to the Crimsonian authorities, as only States can violate sovereignty. The first act of violation of sovereignty deals with the territorial aspect. When a person physically crosses the borders with money and weapons, the involvement of state authorities is more likely than when done virtually by sending money. Berylia might also focus on alleged Crimsonian interference with Berylian governmental functions.

Even though these events are non-cyber in nature, they further fuel the escalation process. If Berylian claims that Crimsonia has violated its sovereignty prove correct, interstate tension will be escalated, and the legal background of this fictional conflict will change. In response, Crimsonia might claim that the deployment of BERM in the disputed region violates Crimsonian sovereignty. As BERM targets any device connecting to specific cell towers near the border, it is possible that the devices of Crimsonian citizens will also be affected. Therefore, it is important to examine BERM capabilities and possible targeting issues before deployment. In our scenario, Berylia should conduct a software review regarding the conditions which Crimsonia may take into consideration when labelling the deployment of BERM as a violation of sovereignty. Violation of sovereignty by cyber means remains an unsettled issue, and the IGE presented three levels which might be helpful to determine whether a violation of territorial sovereignty has occurred. These include considerations as to whether BERM is capable of causing physical damage, loss of functionality or infringement upon territorial integrity falling below the threshold of the loss of functionality.[33] It is also important whether the deployment of BERM leads to interference with the inherently governmental functions of Crimsonia.[34]

30 Island of Palmas Case (Netherland, USA), 4 April 1928, 838.

31 Michael N. Schmitt and Liis Vihul, 'Respect for Sovereignty in Cyberspace', (2017) 95 Texas Law Review 1639; Gary P. Corn and Robert Taylor. 'Sovereignty in the Age of Cyber', (2017) 111 AJIL Unbound 207.

32 Speech by the Attorney General Jeremy Wright at Chatham House delivered on 23 May 2018. < https://www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century> [accessed 17 December 2019].

33 *Tallinn Manual 2.0*, supra n. 6, 20.

34 *Tallinn Manual 2.0*, supra n. 6, 21.

B. Non-Intervention

1) Scenario

With the progression of the Berylian counter-terrorism operation, Crimson Home quickly depletes its human resources and its members are arrested or incapacitated as a direct result of actions by Berylian law enforcement and armed forces. The border, so far used for transportation of cash and weapons, is crossed by people willing to join Crimson Home. Berylian intelligence suggests that these volunteers are affiliated to Crimsonian paramilitary and military forces. However, direct and clear evidence is lacking. BERM is deployed to target devices connecting to specific cell towers located near the border. The payload still effects mainly surveillance of movement and surveillance of communication with intended recipients within Crimsonian territory.

2) Legal Qualification

The principle of non-intervention has very close ties to sovereignty. It is described as 'a corollary of the principle of the sovereign equality of States'.[35] Non-intervention mainly deals with the 'decision-making capacity of a State to formulate policies in relation to its internal and external affairs'.[36] The concept of internal and external affairs is flexible and linked to the notion of *domaine reservé*. The International Court of Justice (ICJ) sheds some light on the definition and has held that States may decide freely on matters such as 'choice of a political, economic, social and cultural system, and the formulation of foreign policy'.[37] That being said, not every coercion trying to violate this freedom of choice violates international law. Only coercive acts reaching a sufficient level of magnitude and intending a target State to change its policy are legally relevant.[38] However, this threshold is fluid and context-dependant.

Berylia might assess whether dozens of people crossing the border and willing to fight for Crimson Home constitute a violation of the principle of non-intervention. Individuals are not legally capable of violating the non-intervention principle. Therefore, Berylia should probably resort to a political attribution and make its suspicion of affiliation of volunteers to Crimsonian forces public. Berylia should also take into consideration the context and intent. Crimson Home sought to secede the region through a referendum and when denied, it turned to violence. *Ergo*, it is pushing for a change of Berylian policy with regard to the disputed region. If the personnel joining Crimson Home intend to force Berylia to change its position towards the region, it might suffice to establish an unlawful intervention.

35 ICJ, Case Concerning Military and Paramilitary Activities in and Against Nicaragua ('*Nicaragua v. United States of America*'), Judgment of 27 June 1986, para. 202.

36 Russell Buchan. 'Cyber Attacks: Unlawful Uses of Force or Prohibited Interventions?' (2012) 17 *Journal of Conflict and Security Law* 212, 223.

37 ICJ, *Nicaragua v. United States of America*, supra n. 35, para. 205.

38 Buchan, supra n. 36, 223-224.

Under international law, Berylia is entitled to engage in countermeasures. As BERM is already deployed, it might serve the purpose besides the collection of data for intelligence, counter-intelligence and law enforcement purposes. It would be necessary to conduct a software review to ascertain whether the use of BERM might further escalate the conflict by exceeding what are permissible countermeasures.

C. Use of Force vs. Armed Attack

1) Scenario

While BERM was previously used mainly to gather intelligence, in response to the violation of its borders Berylia engages in remote destruction of data on devices carried by people crossing the border. This leads to loss of data of many innocent citizens from both Berylia and Crimsonia and large-scale damage to and destruction of property. According to Berylian intelligence, this extreme measure was only partially effective in response to Crimson Home and its affiliates. Crimsonia officially and publicly denounces the deployment of BERM and the harmful payloads distributed through the system. The Crimsonian government also announces that appropriate measures will be undertaken in response. This results in cyber attacks against Berylian dual-use and military infrastructure. Most of these attacks are DDoS and ransomware, but Berylian intelligence reports that they are serving as decoys for large-scale intelligence gathering and espionage. The communications of Berylian forces engaged in ongoing Berylian counter-terrorism operations within the disputed region are jammed from Crimsonian territory.

2) Legal Qualification

Remote destruction of data escalated the situation. We argue that the Berylian action and Crimsonian reaction pushed the whole conflict over the threshold of the use of force, making it inconsistent with purposes enshrined in Article 2(4) of the UN Charter. The IGE partially followed the *scale and effect*s approach laid out by the ICJ,[39] and used this approach for the qualification of the unlawful use of force.[40] To ease the qualification, the IGE also used a set of eight factors[41] that outline factual considerations on whether to consider a given cyber operation as an unlawful use of force. Despite these factors not being norms of international law, they do provide basic cues along which to structure the legal response.[42]

Before using BERM to deploy payload that might lead to a violation of the prohibition of the use of force, Berylia should have conducted a legal review to assess the possible legal consequences to determine, amongst other things, whether the operation may

39 ICJ, *Nicaragua v. United States of America*, supra n. 35, para. 195.
40 *Tallinn Manual 2.0*, supra n. 6, 331.
41 Factors include severity, immediacy, directness, invasiveness, measurability, military character, State involvement and presumptive legality. Compare *Tallinn Manual 2.0*, supra n. 6, 334-336.
42 As emphasised by the IGE 'they are merely factors that influence states making use of force assessments; they are not formal legal criteria'. *Tallinn Manual 2.0*, 333.

lead to violation of Article 2(4) of the UN Charter. Furthermore, Article 51 of the UN Charter which grants a victim state the option to respond with force comes into play. Even though the majority of States perceive the gap between the use of force and an armed attack and distinguish 'the most grave forms of the use of force (those constituting an armed attack) from other less grave forms',[43] other approaches also exist in the international community. As Harold Koh said at the Inter-Agency Legal Conference in 2012, 'the United States has for a long time taken the position that the inherent right of self-defence potentially applies against any illegal use of force'[44] and rejected the existence of any threshold. The other theory called the accumulation of events doctrine was originally introduced by Israel in the 1970s and reflected a situation of terrorist attacks. Israel advocated a position that even though:

> 'each specific act of terrorism, or needle prick, may not qualify as an armed attack that entitles the victim State to respond legitimately with armed force, the totality of the incidents may demonstrate a systematic campaign of minor terrorist activities that does rise to the intolerable level of armed attack.'[45]

D. Armed Conflict

1) Scenario

Berylian intelligence has obtained conclusive proof that volunteers crossing the border from Crimsonia are predominantly members of the Crimsonian armed forces and their activities are being organised by the Crimsonian government. The Berylian government publicly accuse Crimsonia of plans to occupy the disputed region by force. Berylia deploys heavy weaponry to the border region as a follow-up to the counter-terrorism operation against Crimson Home. As part of the preparation for potential conflict, BERM is taken over by the military to ensure coordination of intelligence gathering and targeted incapacitation of devices throughout the disputed region.

Newly-deployed Berylian forces engage volunteers from Crimsonia. As one of the Berylian units engages Crimson Home members and volunteers close to the border, the Crimsonian Air Force attacks the unit. As a follow-up, Crimsonia claims that the military build-up in the disputed region signals a planned invasion by Berylia. The Crimsonian government opts to move units across the border to set up defensive positions on a mountain ridge on Berylian territory. In response, Berylian units engage the Crimsonian Army to prevent it from crossing the border to Berylia.

43 ICJ, Nicaragua v. United States of America, supra n. 35, para. 191.
44 Hongju Koh, Harold. 'International Law in Cyberspace', (2010) *Faculty Scholarship Series* 4854, 7.
45 Norman Menachem Feder, 'Reading the U.N. Charter Connotatively: Toward a New Definition of Armed Attack', (1987) 19 *N.Y.U. J. Int'l L. & Pol*. 395, 415.

2) Legal Qualification

We deem that the last round of escalation leads Berylia and Crimsonia into a state of armed conflict. As a result, the norms of IHL are triggered. Berylia is, as a signatory to the API, obliged to conduct a weapons review under Article 36 of API.

According to the IGE, all States, whether they have ratified API or not, are required to ensure that the means of warfare they acquire or use comply with the rules and principles of IHL. This obligation is derived from a general duty of compliance with IHL.[46] There are at least two points to highlight the weapons review process. First, IHL does not mandate States to establish a general practice of using a weapon before it is to be considered legal.[47] Second, the Commentary to API sheds light on the intent behind the weapons review. It requires States to determine whether the employment of a weapon for its expected use could be prohibited under IHL.[48]

6. DISCUSSION

Although the term 'weapons review' is frequently tossed around, there are different approaches not only between individual States, but also within States themselves. We can take the United States as an example. The United States did not ratify API and its views on reviewing the legality of weapons can be found in the *DoD Law of War Manual* from June 2015 (*the Manual*).[49] It is the position of the DoD to require a legal review for the intended acquisition or procurement of weapons or weapons systems.[50] Such a review should address three questions to determine whether the weapon's acquisition is prohibited with regard to U.S. DoD obligations: (1) whether the weapon's intended use will cause superfluous injury; (2) whether the weapon is inherently indiscriminate; and (3) whether the weapon falls within a type that has been specifically prohibited.[51] U.S. DoD approaches these legal reviews in two stages. The first is an evaluation of the weapon to determine whether its use would be illegal *per se*. The second is to determine whether its use in a particular operation could be illegal.[52] The *Manual* also addresses the legal review of weapons employing cyber capabilities. It notes that not all cyber capabilities constitute a weapon and it is up to individual branches (i.e. Army, Navy, Air Force) of the US armed forces to determine which cyber capabilities require legal review. The *Manual* highlights the most obvious

46 *Tallinn Manual 2.0*, supra n. 6, 464-465. The IGE commented that this duty of compliance is reflected in Article 1 of the 1907 Hague Convention IV and Common Article 1 of the 1949 Geneva Conventions.

47 Office of General Counsel, Department of Defence, *Department of Defence Law of War Manual 338* (2015, updated 2016).

48 Louise Doswald-Beck and Jean-Marie Henckaerts. Customary International Humanitarian Law 237 (2005). The basis for this principle, which reflects customary international law, is Article 23(e) of the Hague Regulations and Article 35(2) of API.

49 *DoD Law of War Manual*, supra n. 46, 337.

50 Ibid.

51 Ibid., 338-9.

52 Ibid., 1025.

IHL related concern, which is the potentially indiscriminate effect of a cyber weapon. It notes that a destructive computer virus designed and intended to spread and destroy uncontrollably within the civilian internet systems and networks would be prohibited under IHL as an inherently indiscriminate weapon.[53]

The term 'weapon' is used in different contexts and often without the normative meaning given to it by international law. The same could be said of the term 'weapons review', as it immediately brings out requirements according to Article 36 of API.

That being said, it is undeniable that the use of software for security purposes has consequences in terms of international law both in State-to-State and State-to-individual relationships. Additionally, individual cyber systems can be used to deliver different payloads, and it is hard to pinpoint the exact moment at which the payload becomes a weapon. A broader understanding of software review concerning international law obligations is sensible. We argue that this sort of review entails practical necessity. The development of new software might be quite costly and the guide to the legal review of new weapons states 'conducting legal reviews at the earliest possible stage is to avoid costly advances in the procurement process (which can take several years)'.[54] This applies even outside armed conflict and the weapons review prescribed in Article 36 of API.

The violation of legal obligations, as our scenarios illustrate, can happen on many different levels in conflict. Article 36 of API prescribes review to prevent the violation of IHL norms. We argue that a system of broader software review would bring (1) more understanding of legal consequences in general, and (2) better framing of policy responses in terms of escalation and de-escalation of potential conflicts.

7. CONCLUSION

In formulating our research question of what legal requirements need to be considered when deploying cyber weapon in situations below the threshold of armed conflict, our broader intent was to evaluate whether the requirement of legal review of cyber weapons or capabilities exists outside IHL. We used a fictional scenario of an escalating conflict, presented basic facts and legal qualification of different events.

The conclusion is, there are plenty of legal requirements to be considered when deploying cyber means. These range from human rights obligations and their possible derogation in case of internal emergency all the way to IHL in armed conflict. Rather unsurprisingly, we conclude that there is no obligation to conduct a review outside Article 36 of API. However, in terms of practical necessity, it is worth considering

53 Ibid., 1025-6.
54 *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare*, supra n. 19, 20.

a broader software review. This would allow more respect to international law obligations by prior evaluation if any software, whether considered a cyber weapon or not, violates the international law obligations of a State.

Cyberspace has brought to light many definitional issues that are still unresolved. A broader approach to software review will allow us to understand the use of software in context and eventually bypass the normative outcomes of labelling something a cyber weapon. We conclude there is no obligation to conduct weapons review outside Article 36 of API. That being said, we believe there are policy benefits in conducting broader software assessments with regard to legal obligations.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

R2P & Cyberspace: Sovereignty as a Responsibility

Tina J. Park, PhD
Vice President,
NATO Association of Canada
& Executive Director and Co-Founder,
Canadian Centre for the Responsibility to Protect
University of Toronto
Toronto, ON, Canada
executive.director@ccr2p.org

Michael Switzer
Deputy Executive Director
Canadian Centre for the Responsibility to Protect
University of Toronto
Toronto, ON, Canada
Michael.M.T.Switzer@gmail.com

Abstract: The Responsibility to Protect, commonly referred to as R2P or RtoP, is an emerging norm in international relations which states that when a state or government fails to protect its people from mass atrocity crimes, the international community has the responsibility to do so. First coined in 2001 and later adopted by 150 heads of state and government at the 2005 World Summit, R2P has been hailed as the most important turning point for the notion of 'sovereignty as responsibility'. Yet, to date, no proper attention has been given to understanding how the technological changes in cyberspace affect the prevention and response to R2P crimes at the national, regional and international levels. This paper explores how evolving cyber capabilities relate to the facilitation, commission and prevention of mass atrocity crimes, specifically war crimes, crimes against humanity, genocide and ethnic cleansing, under the Responsibility to Protect framework in order to (A) demonstrate that such capabilities should be examined and incorporated into the R2P discourse and (B) recommend measures to bolster the efficacy of this incorporation. It begins by discussing the historical significance of R2P, exploring its current conceptual framework and making a case for why prevention efforts deserve consideration. It then proceeds to examine three broad categories in the cyber domain (material sabotage, information collection and social influence) which may be relevant to prevention efforts of R2P. The article concludes with recommendations for more effective integration of cyber capabilities

to prevention efforts and ultimately argues that a greater attention must be given to the relationship between R2P and the cyber domain.

Keywords: *responsibility to protect (R2P), sovereignty as responsibility, prevention, mass atrocities, cyberspace*

1. INTRODUCTION

On 5 November 2018, Facebook admitted that it had failed to prevent its platform from 'being used to foment division and incite offline violence' amid the ongoing ethnic cleansing of the Rohingya people in Myanmar.[1] However, such incitement was not a random incident. It represented part of a campaign, expressed through cyber means, to create the conditions for the mass atrocity that is currently unfolding in Myanmar. As the *New York Times* reported, '[m]embers of the Myanmar military were the prime operatives behind a systematic campaign on Facebook that stretched back half a decade and that targeted the country's mostly Muslim Rohingya minority group'.[2] In fact, while the widespread use of Facebook as a platform for inciting hate may be a recent phenomenon, the use of communications technology in augmenting the commission of mass atrocity crimes is nothing new. For instance, in the build-up to the 1994 Rwandan genocide, Hutu elites used the Radio Mille Collines to incite hatred against Tutsis and Hutu moderates.[3] Once the killings began, the radio was used to relay instructions, lists of names and messages of support to *génocidaires* throughout the country.[4] In turn, the Rwandan genocide saw the most efficient and ruthless massacre of some 800,000 innocent lives over the course of merely a hundred days, while the international community remained as silent bystanders.[5]

The cases of Myanmar and Rwanda both demonstrate a well-known fact: mass atrocity crimes do not happen overnight and technology can be easily used or abused for these incidents. With proper early warning systems and efficient response mechanisms in the cyber domain, they can be prevented and halted in a timely manner. These crimes present a clear shock to values codified in the Universal Declaration of Human Rights

1 Alex Warofka, 'An Independent Assessment of the Human Rights Impact of Facebook in Myanmar,' *About Facebook*, November 5, 2018, https://about.fb.com/news/2018/11/myanmar-hria/.

2 Paul Mozur, 'A Genocide Incited on Facebook, With Posts From Myanmar's Military,' *The New York Times*, October 15, 2018, https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.

3 Evaina Bonnier, Jonas Poulsen, Thorsten Rogall, Miri Stryjan, 'Preparing for Genocide: Quasi-Experimental Evidence from Rwanda'. (No 31, SITE Working Paper Series from Stockholm School of Economics, Stockholm Institute of Transition Economics, 2015), 25. https://pdfs.semanticscholar.org/d78f/0bc73c715b9af13d69f9afeaedc4cbfb30bd.pdf.

4 Samantha Power, *A Problem from Hell: America and the Age of Genocide*, (New York: Basic Books, 2013), 371.

5 Power (supra n. 4), 327.

and to our collective conscience. The pledge of 'never again' has been enshrined in the Responsibility to Protect (R2P), an international norm which asserts that when sovereign states are unable or unwilling to fulfil their responsibility to protect their own populations from mass atrocity crimes, the international community has a responsibility to do so. While the military intervention aspect of R2P has been quite controversial since the inception of the principle, R2P still represents an important milestone in perceiving sovereignty as responsibility. One of R2P's primary strengths lies in its holistic approach to prevention. In promoting prevention as a key pillar of R2P, the UN Office on Genocide Prevention and the Responsibility to Protect has created a Framework of Analysis for Atrocity Crime Prevention (referred to henceforth as 'the Framework'), which helps to identify the risks for the commission of mass atrocity crimes and produces a series of indicators to guide prevention efforts.[6] In this way, R2P's operational framework focuses not only on how we may address atrocity crimes, but also on the factors that give rise to such crimes.

This article analyses the relationship between the evolution of the cyber domain and the prevention of R2P crimes – and, more importantly, how the international community may best leverage cyber capabilities to advance R2P's ultimate objective: a world free from mass atrocity crimes.

This article will underline the following key arguments: first, cyber capabilities should be incorporated into efforts to implement R2P; second, the application of R2P must be widened to include private sector partnership, especially in the prevention stage. This article is divided into three parts. First, it examines R2P's historical significance, theory and preventative utility. Second, it argues that there are already key points of intersection between cyber capabilities and R2P – which presents both challenges to and opportunities for prevention. Lastly, it argues, on the basis of this examination, for the incorporation of cyber capabilities into R2P – concluding with suggestions for moving forward.

2. WHAT IS R2P? EVOLUTION OF R2P FROM 2001 UNTIL THE PRESENT

A. Origins of R2P: The 2001 ICISS Report

To understand how R2P is different from the classic conception of humanitarian intervention, it is useful to examine the norm's origins. Following the end of the Cold War, the rise of conflicts in Bosnia and Herzegovina, Somalia, Rwanda and Kosovo gave rise to the notion of humanitarian intervention.[7] This concept proved

6 United Nations, *Early Warnings*, United Nations Office on Genocide Prevention and the Responsibility to Protect, https://www.un.org/en/genocideprevention/early-warning.shtml.

7 Jennifer Welsh, 'Authorizing humanitarian intervention,' in Richard M. Price and Mark W. Zacher eds., *The United Nations and Global Security*, (Basingstoke: Palgrave Macmillan, 2004), 177– 192.

highly controversial. When interventions were undertaken in Somalia, Bosnia and Kosovo, they were heavily criticised.[8] However, when interventions failed to take place – particularly in the case of Rwanda – such inaction came with unfathomable human costs.[9] Following debates over the unilateral NATO intervention in Kosovo in 1999, UN Secretary-General (UNSG) Kofi Annan urged UN member states to 'find a common ground' in upholding the principles of the Charter and acting in defence of our common humanity.[10] In response, the government of Canada sponsored the creation of the International Commission on Intervention and State Sovereignty (ICISS), which released its report, *The Responsibility to Protect*, in 2001.[11]

R2P, as advanced by the Commission, consisted of three key responsibilities with respect to the protection of populations: a responsibility to prevent situations in which such harm could occur; a responsibility to react to such situations; and a responsibility to rebuild following their conclusion.[12] The ICISS report marked two notable conceptual shifts. The first was a recognition that, to reconcile the debate between non-intervention and humanitarianism, it was necessary to see a state's sovereignty as implying a responsibility to protect its own population.[13] The second was a shift in the conceptual language surrounding the response to humanitarian disasters. This encompassed a change from the language of 'humanitarian intervention' which focused on the rights of intervening states, to the language of a 'responsibility to protect,' which focused on the state's duty to protect its population.[14]

B. Adoption of R2P: 2005 World Summit Outcome Document and SG's annual report on R2P

The R2P advanced by the ICISS report did not immediately take effect on the international stage, especially as the international community became occupied by the Sept 11 attacks and the 'War on Terror'. From 2001 onward, a group of 'norm entrepreneurs' came together to promote its mainstream acceptance by UN member states.[15] Their efforts met with significant success in 2005, when R2P was adopted in paragraphs 138 and 139 of the UN World Summit Outcome Document (WSOD). These paragraphs were important to the development of R2P in three respects. First,

8 International Commission on Intervention and State Sovereignty (ICISS), *The Responsibility to Protect*, (Ottawa: International Development Research Centre, 2001), Introduction.

9 Ibid.

10 United Nations Report of the Secretary General, *In Larger Freedom: Towards Development, Security and Human Rights for All*, A/59/2005, (2005), paragraph 220, https://undocs.org/A/59/2005.

11 Brian Tomlin, Norman Hillmer and Fen Hampson, *Canada's International Policies: Agendas, Alternatives and Politics*, (Toronto: Oxford University Press, 2008), 214-215.

12 International Commission on Intervention and State Sovereignty (ICISS), *The Responsibility to Protect*, (Ottawa: International Development Research Centre, 2001), xi.

13 Gareth Evans, *The Responsibility to Protect: Ending Mass Atrocity Crimes Once and For All* (Washington, DC: Brookings Institution Press, 2008), 43.

14 Ramesh Thakur, "R2P After Libya and Syria: Engaging Emerging Powers." *The Washington Quarterly* 36, no. 2 (2013), 65.

15 Tina J. Park and Victor MacDiarmid. "Selling R2P: Time For Action." In John Forrer and Conor Seyle eds., *The Role of Business in the Responsibility to Protect*, (Cambridge: Cambridge University Press, 2016), 1.

the fact that their adoption was unanimous demonstrated an international consensus on the norm. Second, great care was taken to the final articulation of R2P: the wording of paragraphs 138 and 139 were results of intense debates involving perspectives from a diversity of regions. Third, the version of R2P they advanced was different from that of the ICISS – largely due to the requirements of unanimity and compromise. While the WSOD's R2P advanced the norm's focus by constraining its scope to the four mass atrocity crimes, none of the ICISS's six criteria concerning intervention were included, nor was there any mention of a responsibility to rebuild.[16]

C. R2P Today: The Norm's Three Pillars

Since 2005, R2P has evolved as an international norm that draws on existing international law to define the responsibilities of states and the international community regarding four narrowly-defined mass atrocity crimes. R2P is an international norm in that it does not add legal obligations that constrain or determine behaviour; instead, as any norm does, it advances a shared standard of appropriate action for states, international organisations, civil society and private sector entities.[17] R2P's normative evolution is best reflected in former UN Secretary-General Ban Ki-moon's 'Three Pillar' framework. Articulated in his 2009 UN report entitled *Implementing the Responsibility to Protect*, this framework translates the commitment to R2P expressed by paragraphs 138 and 139 in the World Summit Outcome to the following three responsibilities, which are to be employed simultaneously:[18]

- **Pillar One:** Individual states have a responsibility to protect their populations from the commission and incitement of genocide, war crimes, ethnic cleansing and crimes against humanity.
- **Pillar Two:** The international community, member states, civil society and the private sector are responsible for assisting individual states in meeting their pillar one responsibilities – particularly in the context of preventing mass atrocity crimes.
- **Pillar Three:** UN member states have a responsibility to 'respond collectively in a timely and decisive manner' when a member state fails its pillar one obligations. This response must be in accordance with the 'provisions, principles and purposes' of the UN charter; while such a response could include the use of force, this measure can only be legitimate when it is authorised by the UN Security Council.[19]

These pillars form the basis of the modern conceptual understanding of R2P and are important in two aspects. First, the framework highlights a multitude of proactive

16 Yaroslav Radziwill, *Cyber-Attacks and the Exploitable Imperfections of International Law*, (Leiden: Brill Nijhoff, 2015), 288.

17 Melissa Labonte, 'R2P's Status as a Norm' in Alex J. Bellamy and Tim Dunne eds, *The Oxford Handbook of the Responsibility to Protect*, (Oxford: Oxford University Press, 2016), 137.

18 United Nations, 2009 Report of the UN Secretary-General, *Implementing the Responsibility to Protect*, A/63/677, (January 12, 2009), paragraph 12, https://undocs.org/A/63/677.

19 Ibid.

measures beyond military intervention to protect populations from atrocity crimes. The prospect of a conventional military response to the commission of atrocity crimes represents a small (albeit important) minority of the actions that R2P advocates, even in its third pillar. Alongside effective reaction, R2P prioritises a wide range of economic and diplomatic prevention methods. As such, a key strength of the R2P framework lies with the fact that it advances a set of actions that focus on ameliorating the root causes of mass atrocity crimes.

Second, the framework is 'narrow but deep' in its scope of only four, well-defined crimes.[20] Genocide, war crimes and crimes against humanity have explicit definitions in existing pieces of international law,[21] while existing *opinio juris* states that the practices that define ethnic cleansing can be assimilated into these crimes.[22] While this approach may lead to issues of contestation over applying the definition of these crimes to real-world examples, the narrowing of this scope ensures that consensus about the principle endures.[23]

D. R2P's Current Status Post-Libya: Holistic Prevention

In view of the controversial implementation of UNSC Resolution 1973 in Libya and subsequent P5 deadlock in Syria, R2P's current focus lies squarely on the strength of its holistic approach to prevention. Libya represented the first public test of R2P's implementation concerning the use of force. Before bestowing the mandate authorising NATO to use 'all necessary means to protect civilians' in resolution 1973,[24] the UN exhausted 'eleven out of the thirteen' alternative measures for which R2P advocates, including economic sanctions, preventative military deployment and arms embargoes.[25] However, as the intervention progressed, coalition leaders argued that a 'real and lasting protection of civilians could not take place with Qadhafi in power' and hence, he must be deposed.[26] This interpretation proved controversial, drawing sharp criticism from Brazil, China, India, Russia and South Africa, who charged the mission with overstepping its mandate.[27]

20 Jennifer Welsh, 'The 'Narrow but Deep Approach' to Implementing the Responsibility to Protect: Reassessing the Focus on International Crimes,' in Rosenberg, Sheri P., Tibi Galis, and Alex Zucker eds., *Reconstructing Atrocity Prevention*, (Cambridge: Cambridge University Press, 2015), 82.

21 United Nations, *Framework of Analysis*, https://www.un.org/en/genocideprevention/early-warning.shtml Annex I. Genocide is defined in Article 2 of the Convention on the Crime of Genocide; Crimes against humanity are defined in article 7 of the Rome Statute; and War Crimes are defined in article 8 of the Rome Statute.

22 United Nations, *Framework of Analysis*, 32.

23 Jennifer Welsh, 'The 'Narrow but Deep Approach' to Implementing the Responsibility to Protect: Reassessing the Focus on International Crimes,' *Op.Cit*.

24 UN Security Council Resolution 1973 (2011), S/RES/1973 (17 March 2011), https://www.undocs.org/S/RES/1973%20.

25 Paul Tang Abomo, *R2P and the US Intervention in Libya*, (New York; Palgrave Macmillan, 2018), 243.

26 Barack Obama, David Cameron and Nicolas Sarkozy, 'Libya's Pathway to Peace', *The New York Times*, 14 April 2011.

27 Alex J. Bellamy and Tim Dunne eds. *The Oxford Handbook of the Responsibility to Protect*, (Oxford: Oxford University Press, 2016), introduction.

Nevertheless, these criticisms do not translate into an outright rejection of R2P, nor do they negate the incredible degree of progress made with this emerging norm in the past few decades. Rather, the case of Libya served as a test for whether the international community could react to mass atrocity crimes in a way that solely concerned the protection of populations. As a result, little international action has been taken to stem ongoing atrocities committed by government forces in Syria. Rather, China and Russia are primarily concerned about R2P being used as a tool for regime change.[28] In this way, the 2011 intervention in Libya has merely precluded the military application of R2P's third pillar; R2P's second-pillar suite of non-military preventative measures – ranging from fostering economic stability and combating hateful ideologies to ensuring transparency in criminal justice systems – do not allow the same possibility for regime change. As such, R2P's prevention measures are far less rigid in the forms they may take, allowing for actors to find common ground, with excellent opportunities for cyber activities.

3. CYBER DOMAIN AND R2P: KEY POINTS OF RELEVANCE

To assess challenges and opportunities regarding R2P and cyber domain, this section will begin with a definition of these categories: Cyber Material Sabotage (cMS), Cyber Information Collection (cIC) and Cyber Social Influence (cSI). It will then define the ways in which each category of capability represents challenges or opportunities relevant to R2P. Cyber capabilities are defined not as technologies, but rather as the potential for an actor to effect change through a particular channel of technology in the cyber domain. This section will draw upon the UN Framework of Analysis for Mass Atrocity Prevention, to see how cyber capabilities can have implications on mass atrocity crimes.

A. Definitions: Three Categories of Cyber Capability

- **Cyber Material Sabotage (cMS)** capabilities enable an actor to damage another actor's capacity to function.
- **Cyber Information Collection & Manipulation (cICM)** capabilities enable an actor to obtain, organise and manipulate information about a population, institution, agency or operation – albeit in a way that does not cause material damage.
- **Cyber Social Influence (cSI)** capabilities enable an actor to alter the perceptions, beliefs and decision-making of a given population.

28 Thakur (supra n. 14), 71.

These three categories are not meant to exhaust the range of possibilities that an actor may realise through technological tools in the cyber domain. Instead, they are designed to allow for a more effective discussion of those sets of possibilities that are most relevant to the R2P framework.

B. Cyber Material Sabotage (cMS) Capabilities

1) Challenges

The Cyber Material Sabotage (cMS) capabilities present challenges to R2P in two ways. First is the tangible damage that cyber operations targeting financial or institutional infrastructure can cause with regards to the stability and resilience of a society. Such disruptions could constitute measures that fall under indicator 8.9 of the UN *Framework of Analysis*, namely: 'Sudden changes that affect the economy or the workforce, including as a result of financial crises, natural disasters, or epidemics'.[29] For example, in the event that any cMS capability is used to target financial services or infrastructure, it may have far-reaching consequences that could seriously disrupt the quality of life and economic stability. Examples of these disruptions are the hacking that crippled South Korean banks and infrastructure, including Korea Hydro, or the cyber-attacks on Sony and some American banks.[30]

Second, cyber thefts represent not only harm to a particular organisation or economy's capacity to function, but also an ever-growing stream of revenue that can bolster the capacities of actors to commit mass atrocities, especially collaborations with non-state terror groups. For example, since it began its cyber operations, North Korea has reportedly acquired as much as $USD 2 billion through cyber activities.[31] Much of these funds have also been successfully laundered online.[32] This revenue, in turn, has been used to develop weapons, ranging from nuclear weapons to chemical and biological weapons, which the North Korean regime sells to non-state terror groups in the Middle East such as Hezbollah and Hamas. Because of the very nature of cyber crimes and difficulties with attribution, the cMS capabilities pose real and serious threats to regional and international security, as well as day to day lives of ordinary citizens.

29 United Nations. *Framework of Analysis for Atrocity Crimes: A Tool for Prevention*. (2014). https://www.un.org/en/genocideprevention/documents/about-us/Doc.3_Framework%20of%20Analysis%20for%20Atrocity%20Crimes_EN.pdf.

30 Mattha Busby, 'North Korean 'Hacker' Charged over Cyber-Attacks against NHS,' *The Guardian* (Guardian News and Media, September 6, 2018), https://www.theguardian.com/world/2018/sep/06/us-doj-north-korea-sony-hackers-chares.

31 Edith M. Lederer, 'UN Probing 35 North Korean Cyber Attacks in 17 Countries,' *Associated Press*, (August 13, 2019), https://apnews.com/ece1c6b122224bd9ac5e4cbd0c1e1d80.

32 Michelle Nichols, 'North Korea Took $2 Billion in Cyberattacks to Fund Weapons Program: U.N. Report,' *Reuters* (August 5, 2019), https://www.reuters.com/article/us-northkorea-cyber-un/north-korea-took-2-billion-in-cyberattacks-to-fund-weapons-program-u-n-report-idUSKCN1UV1ZX.

2) Opportunities

The cMS capabilities present a limited set of opportunities for acceptable use as preventative tools. For instance, it would be illegal for states to use cMS capabilities against other states unless authorised by the UNSC, which, in the wake of Libya, may be unlikely. The *Tallinn Manual* makes it clear that: 'A State may not intervene, including by cyber means, in the internal or external affairs of another State'.[33] This encompasses (1) situations in which states intervene through cyber means and (2) situations in which states intervene in the cyber affairs of another state using non-cyber coercive means.[34] In either case, the *Tallinn Manual* asserts that 'a prohibited act of intervention' requires that 'the act in question must relate to matters that involve the internal or external affairs of the target State' and that the act 'must be coercive in nature'.[35] As cMS capabilities are, by their disruptive nature, inescapably coercive, it is unlikely that the cMS capabilities may be legitimately used by states against states.

However, there is already precedent for the use of cMS capabilities against non-state actors – for example, in 2016, the US military conducted its first offensive cyber operation against ISIS with the aim of disrupting the organisation's finances, recruiting and propaganda.[36] By reducing the financial and logistical capacity of potential non-state perpetrators, such measures take a preventative approach towards promoting the rights and freedoms of ordinary citizens. Furthermore, building a strong defence against these cMS capabilities at the national level could help foster the resilience of any society, long before any mass atrocity crimes take place.

C. Cyber Information Collection & Manipulation (cICM) Capabilities

1) Challenges

The Cyber Information Collection and Manipulation (cICM) capabilities pose challenges to R2P in two ways. First, surveillance capabilities engendered by facial recognition, GPS-tracking and the access to data transmitted through information and communication technologies (ICTs) allows actors to identify and track populations based on certain attributes, which relates to indicator 7.12 of the UN's *Framework of Analysis* by bolstering their capability to 'mark people or their property based on affiliation to a group'.[37] Indeed, the November 2019 leak of four classified Chinese bulletins illustrates the ways in which China has combined a variety of surveillance capabilities to create the Integrated Joint Operations Platform (IJOP), which included 'a detailed database of everything from an individual's exact height and electricity

33 Michael N. Schmitt et al., *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*, (Cambridge: Cambridge University Press, 2017), 313.

34 Ibid.

35 Ibid. 314.

36 Ash Carter, 'A Lastine Defeat: The Campaign to Destroy ISIS,' *Harvard Belfer Center Report*, (October 2017), 32. https://www.belfercenter.org/sites/default/files/2017-10/Lasting%20Defeat%20-%20final_0.pdf.

37 United Nations, Report of the Secretary General, *Responsibility to Protect: Lessons learned for Prevention*, A/73/2019, (2019), paragraph 23, https://digitallibrary.un.org/record/3810380?ln=en.

use, to the colour of their car,' to 'if they prefer to use the front or back door to their house'. The capacity conferred by the IJOP is substantial enough that it flagged 24,412 suspicious individuals, 'of which 15,000 were sent to re-education camps and a further 706 were jailed'.[38]

Second, cyber capabilities allow for a greater degree of control over the amount and kind of information in circulation within certain sectors of cyberspace. Such cICM capabilities allow for the suppression of information relating to the early identification of mass atrocity crimes. For instance, China's 'control of court data' and 'media censorship of cases' contributed to the difficulty in assessing the extent of China's detention of journalists.[39] The capability for an actor to hide the extent of its persecution complicates prevention efforts in the way of indicator 6.11 in the UN *Framework of Analysis*, namely the 'lack of an early warning mechanism relevant to the prevention of atrocity crimes'.[40] Lastly, the emergence of 'deep fakes', such as videos generated via algorithms that make it look like a person said or did something she did not, allows actors to tamper with video evidence so as to avoid accountability, relating to indicator 3.6 of the UN *Framework of Analysis* specifically: 'Absence or inadequate external or internal mechanisms of oversight and accountability'[41] or alter a population's perception of reality through propaganda.[42]

2) Opportunities

On opportunities, the first promising cICM capability is the ability to record and monitor security forces. This capability plays a supportive role in bolstering accountability and the rule of law by providing a more transparent method of monitoring police and security forces. An example of this may be found in the creation and storage of police footage. A 2018 article by the US National Institute of Justice notes that the use of 'body-worn cameras' (BWC) by police forces may bolster transparency, allow for the storage of footage to be used as evidence in court proceedings and ensure greater capacity to refine training methods and operational strategies.[43]

All these effects may bolster the state's accountability in upholding the rule of law: knowing that police interactions are recorded may bolster the trust that the public feels towards police forces; storing police footage allows for a better capacity to hold officers accountable for their actions in a court of law; and using footage to refine

38 Emma Graham-Harrison and Juliette Garside, 'Revealed: Power and Reach of China's Surveillance Dragnet,' *The Guardian* (November 24, 2019), https://www.theguardian.com/world/2019/nov/24/china-cables-revealed-power-and-reach-of-chinas-surveillance-dragnet).

39 US Congress, Congressional-Executive Commission on China, *Annual Report 2019*, 116th Cong., 1st sess., 2019. S. Exec. Rep. 36-743, 42, https://www.cecc.gov/publications/annual-reports/2019-annual-report.

40 United Nations, *Framework of Analysis*, 15.

41 Ibid. 12.

42 Ibid. 15-16.

43 Brett Chapman, 'Body-Worn Cameras: What the Evidence Tells Us,' National Institute of Justice, Nov 14, 2018, https://nij.ojp.gov/topics/articles/body-worn-cameras-what-evidence-tells-us.

police methods may boost capacity to improve police-public relations. Indeed, even acknowledging that the BWC are present may make a positive difference. The 2016 'global, multisite randomised controlled trial'[44] study by Ariel et al. found that BWC "can reduce police use of force...when officers' discretion to turn cameras on or off is minimised".[45] Decreasing the prevalence of the use of force by police officers may boost relations between the police and the public, making it more difficult for them to be leveraged as instruments for atrocity crimes.

Second, cICM capabilities allow civilians and journalists, through smartphones and ICTs, to collect and organise media which documents risk factors for mass atrocity crimes. This is useful for effective prevention in three ways. First, ICTs can bolster prevention efforts by serving as conduits for early warning and mobilisation. For instance, during the Egyptian Revolution, protestors circumvented state censorship of the media by using smartphones to document instances of police brutality and political repression, spreading this information to international audiences and providing a wealth of evidence behind which the international community rallied.[46] Second, the ability of those undergoing active atrocity crimes to self-report enables such actors to supply a constant stream of information to policy-makers and the wider public.[47] For example, the recent leak of 24 documents relating to the Chinese internment of Muslim populations in Xinjiang has sustained broad international interest in actions that may constitute crimes against humanity.[48] Third, timely information produced by the use of digital equipment has created a new and fruitful body of potential evidence.[49] This is already the case, as both of the ICC warrants for Libyan Commander Al-Werfalli relied on videos drawn from social media.[50]

Third, while ICTs allow hate speech to propagate with ease, such speech is subject to tools of quantification and analysis. Such tools have already been applied. For example, Mondal et al. undertook a systematic measurement and analysis of hate speech on social media in 2017, allowing them to map the prevalence, targets and geographical distribution of such speech.[51] Online initiatives that have capitalised on this opportunity already exist, such as Hatebase, Islamophobic incident reporting

44 Ibid.
45 Ibid.
46 Ibid.
47 Rebecca Hamilton, 'Atrocity Prevention in the new media landscape,' AJIL Unbound, 113 (2019), 266.
48 Austin Ramzy and Chris Buckley, ''Absolutely No Mercy': Leaked Files Expose How China Organized Mass Detentions of Muslims,' *The New York Times* (The New York Times, November 16, 2019), https://www.nytimes.com/interactive/2019/11/16/world/asia/china-xinjiang-documents.html).
49 Lindsay Freeman, 'Digital Evidence and War Crimes,' *Fordham International Law Journal*, vol 41, issue 2 (2017).
50 Alexa Koenig, ''Half the Truth Is Often a Great Lie': Deep Fakes, Open Source Information and International Criminal Law,' AJIL Unbound 113 (2019): 250-255, https://doi.org/10.1017/aju.2019.47), 251.
51 Mainack Mondal et al. 'A Measurement Study of Hate Speech in Social Media,' HT '17: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, July 2017, https://homepages.dcc.ufmg.br/~fabricio/download/HT2017-hatespeech.pdf, 10.

platforms and Fight against Hate. Such capacity for observation, measurement and analysis is significant for two of the UNSG's preventative recommendations and can bolster state efforts to target hate speech. As Mondal et al. observed, their data 'might provide a unique opportunity to identify the root causes' of 'offline hate'.

Understanding the bigger picture of online hate allows for states to understand and therefore target the root causes of hatred specific to their context and thus create a more vibrant civil society, as more accurate measurement better informs action. The tools of analysis employed by Mondal et al. may also be employed to measure other indicators of respect for diversity and vibrancy of civil society. Using analytic tools to assess the field of expressed attitudes on social media may open up the possibility for governments to better understand their particular context.

D. Cyber Social Influence (cSI) Capabilities

1) Challenges

The Cyber Social Influence (cSI) capabilities represent challenges to R2P in three ways. First, they allow actors to weaken popular trust in institutions. Second, they allow actors to incite hatred and violence towards a particular group. Third, they bolster the ability for organisations committing mass atrocity crimes to recruit others to their cause, thereby enhancing their capacity to commit atrocity crimes and weaken the political will to react.

First, cSI capabilities represent a challenge to R2P by giving actors the ability to undermine the credibility of institutions, a key contributor to the outbreak of mass atrocity crimes. In that regard, influence campaigns can cast doubt on the fairness of an election, which may constitute a triggering factor under the UN *Framework of Analysis*' indicator 8.8: 'Census, elections, pivotal activities related to those processes, or measures that destabilize them'.[52] As Brangetto and Veenendaal noted, Cyber-Berkut's manipulation of voting data in Ukraine's 2014 election, while in no way influencing the election's outcome, nonetheless weakened trust in the 'credibility of the Ukrainian government in overseeing a fair election process'.[53] There is, therefore, a very important correlation between cSI capabilities and public's trust, long before any major crisis breaks out.

Second, cSI capabilities empower an actor to recruit others to their cause. This can be seen primarily in the case of online radicalisation, in which individuals are persuaded in the cyber domain to serve as an asset or an agent for a particular actor. This process has been employed by both state and non-state actors. Moreover, the

52 United Nations, *Framework of Analysis*, 17.

53 Pascal Brangetto and Matthijs Veenendaal, "Influence Cyber Operations: The Use of Cyberattacks in Support of Influence Operations", in N. Pissanidis, H. Rõigas and M. Veenendaal eds., *The 8th International Conference on Cyber Conflict*, (Tallinn: NATO CCD COE Publications, 2016), 121.

definitions of 'recruitment' and 'asset' can have different meanings depending on the situation, ranging from US citizens unwittingly cooperating with Russian Internet Research agents to organise rallies in the build-up to the US election[54] to German citizens moving to Syria to fight for ISIL.[55] Depending on the kind of recruitment, these capabilities relate to indicators such as the UN *Framework of Analysis*' indicator 7.14: 'Increased inflammatory rhetoric, propaganda campaigns or hate speech targeting protected groups, populations, or individuals' and indicator 5.3: 'Capacity to encourage or recruit large numbers of supporters from populations or groups, and availability of the means to mobilize them'.[56]

Third, cSI capabilities can be used for the incitement of hatred towards a particular group, which speaks to the UN *Framework of Analysis*' indicator 7.14: 'Increased inflammatory rhetoric, propaganda campaigns or hate speech targeting protected groups, populations or individuals' and, depending on the severity of such incitement, indicator 8.7: 'Acts of incitement or hate propaganda targeting particular groups or individuals'.[57] The effects of such influence campaigns can be seen specifically with the Myanmar military's campaign of inciting hatred against the Rohingya populations within that country's borders. As the 2018 report of the Fact Finding Commission of the Office of the High Commissioner for Human Rights concluded, hate speech 'contributed to increased tension and a climate in which individuals and groups may become more receptive to incitement'.[58] Moreover, cyber influence campaigns are highly relevant in inciting violence in general. In the case of German right-wing hate media, 'right-wing anti-refugee sentiment on Facebook predicts violent crimes against refugees in otherwise similar municipalities with higher social media usage', while violence dropped appreciably when internet access went down.[59]

2) Opportunities

These cSI capabilities present opportunities by allowing for measures which attack hate speech. Targeting hate speech includes the censorship of hateful actors through attacks on their presence in cyberspace and proactive positive engagement with their target audience. Examples of censorship include the deletion of hateful social media accounts and pages, as has been tried in the case of removing anti-Muslim

54 US House of Representatives Permanent Select Committee on Intelligence, 'Exposing Russia's Effort to Sow Discord Online: The Internet Research Agency and Advertisements,' https://intelligence.house.gov/social-media-content/.

55 Pilar Cebrian, 'They Left to Join ISIS. Now Europe Is Leaving Their Citizens to Die in Iraq'. *Foreign Policy*, Sept 15, 2019. https://foreignpolicy.com/2019/09/15/they-left-to-join-isis-now-europe-is-leaving-their-citizens-to-die-in-iraq/.

56 United Nations, *Framework of Analysis*, 14.

57 Ibid., 16-17.

58 UN Human Rights Council, Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar, UN Doc. A/HRC/39/CRP.2 (Sept. 17, 2018), para. 1354.

59 Karsten Mueller and Carlo Schwarz. "Fanning the Flames of Hate: Social Media and Hate Crime," *SSRN Electronic Journal*, 2017, 1.

Facebook pages and accounts in Myanmar,[60] ISIS-affiliated twitter accounts,[61] and neo-Nazi websites.[62] These measures may be useful in short-term situations where a constrained number of actors are using online platforms to spark and coordinate violence. While censorship of media runs into a number of operational problems,[63] innovative solutions such as social media councils have been created to facilitate more proactive action against hate speech and the incitement of violence. Such councils can be useful at the early stages of a campaign of hate incitement, where populations have yet to internalise hateful messages and are therefore more open to changes in world-view. Coordination between the offices of the UN, member states and social media platforms would bolster the suppression of hateful messages and the proliferation of anti-hate campaigns. Yet, even if these measures were to be successful, they would not be effective against already-internalised hatred.

4. CONCLUSION: LOOKING FORWARD

For too long, the international community has failed in giving sufficient attention to the importance of the cyber domain in the prevention of and response to mass atrocity crimes. Yet, as this paper has demonstrated, there are many points of relevance when we consider the ways in which technology can harm or strengthen our ability to protect populations in peril. In conclusion, this article calls for the following measures.

First, a stronger partnership with the private sector, such as Google and Facebook, is a necessary first step in increasing our protection capabilities in the cyber domain. With the Artificial Intelligence (AI) revolution, it can be expected that communication within societies will increasingly rely on ICTs. In turn, companies that deliver ICT services are uniquely placed to detect and analyse warning signs, proactively remove content which incites violence and bolster international efforts to counteract the spread of hatred, whether it is through altered media, fake headlines, or inflammatory rhetoric. As such, the implementation of R2P – and, perhaps, its conceptual development – must feature buy-in from the private sector and a long-term collaboration in line of existing tenets of international law.

Second, cyber capabilities alone are not by themselves sufficient tools to prevent or halt mass atrocities; they must be combined with political leadership, existing institutions and financial, legal and social resources within a society. For example, the

60 Hannah Ellis-Petersen, 'Facebook removes accounts associated with Myanmar military,' *The Guardian*, Aug 27, 2018, https://www.theguardian.com/technology/2018/aug/27/facebook-removes-accounts-myanmar-military-un-report-genocide-rohingya.

61 'Twitter takes down 300,000 terror accounts as AI tools improve' *Financial Times*, Sept 19, 2017 https://www.ft.com/content/198b5258-9d3e-11e7-8cd4-932067fbf946.

62 https://www.telegraph.co.uk/technology/2017/08/29/worlds-oldest-neo-nazi-website-stormfront-shut/.

63 Emma Irving, 'Suppressing Atrocity Speech on Social Media,' *American Society of International Law*, (Volume 113, 2019), 260.

rapid collection of evidence – through smartphone cameras or BWC – requires that proper ethical and legal accountability measures are in place. The uses and abuses of data collected through advances in communications technology will depend on our ability to ensure that suitable protectionary measures are undertaken in the fine line between an individual's privacy and the protection of society as a whole.

Third, a long-term strategy must be devised to cope with the demands of the AI revolution in cyberspace and its impact on human rights discourse. In the near future, robots and drones could become perpetrators of crimes covered under the R2P framework, which will blur the boundaries of criminality. Unless we are able to keep up with the pace of changes brought about by the AI revolution, our pledge of 'never again' will remain a hollow promise.

Unfortunately, R2P is all-too-often dismissed as a tool for military intervention or a challenge to state sovereignty. However, at the very core of R2P is the notion of sovereignty as a responsibility. As this article has illustrated, it is important for R2P advocates and the international community at large to realise the potential that lies in proactively engaging the tools from the cyber domain. There may never be a clear blueprint for how best to prevent another genocide. Nevertheless, we all share a collective responsibility to adapt to new realities and seize new opportunities from the cyber domain.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

The Past, Present, and Future of Russia's Cyber Strategy and Forces

Bilyana Lilly
Policy Researcher
Pardee RAND Graduate School
RAND Corporation
Santa Monica, CA, United States
blilly@rand.org

Joe Cheravitch
Defense Analyst
Defense and Political Sciences
RAND Corporation
Arlington, VA, United States
jcheravi@rand.org

Abstract: Russian cyberattacks against military and civilian infrastructure in the West have become a persistent challenge. Despite the importance of this topic and the excellent scholarship already published on these issues, there is a need for more detailed data and analysis on the role of cyberattacks in Russia's security strategy and its reflection in the evolution of Russia's cyber forces. A better understanding of Russia's strategy and cyber actors, particularly the growing role of the military in these issues, can facilitate an improvement in Western governments' policies to defend against future Russian activity. To address this issue, this article will outline the role of information and cyber operations in Russia's information warfare doctrine and will analyze the recruitment efforts and modus operandi of Russia's cyber departments, particularly psychological and cyber operations units within military intelligence. The paper will conclude by examining the likely future of Russia's behavior in cyberspace and how various state-sponsored actors might influence it. The paper asserts that although Russia's doctrine suggests a defensive and cooperative posture in response to threats in the information space, officials' promulgations and military literature reveal a predilection for the development of offensive cyber capabilities and operations, which are shaped by Russia's threat perceptions and doctrine, and the institutional cultures of the departments within the military conducting them.

Keywords: *Russia, cyber, cyber strategy, information warfare, information operations, FSB, GRU*

This paper represents independent work by the authors and does not represent the views of the RAND Corporation.

1. INTRODUCTION

Cyber operations attributed to Moscow are not conducted in a strategic vacuum. They are enabled and shaped by broader geopolitical considerations and the institutional culture of Russia's military, intelligence, and political leadership, as well as by Moscow's evolving approach to asymmetric interstate competition that falls short of all-out conflict. To understand the motivations behind and the constraints of Russia's use of cyber and information operations against perceived adversaries, decision-makers must thoroughly study existing policy and doctrine, particularly its evolution from the immediate post-Soviet period until now, while at the same time striving to attain a more sophisticated comprehension of the actors responsible for executing cyberattacks and digital influence campaigns. This involves research into Russian publications and official documents and more nuanced and updated investigations into the actors behind these efforts, which is now possible in the wake of key Russian campaigns, such as the 2016 effort to undermine the U.S. presidential election, that have generated an unprecedented amount of public information on specific units and personalities. Such investigations can help gird the international community against future operations, while assisting policymakers in determining the viability and course of cyber diplomacy and deterrence.

This article aims to show that there is more continuity than contrast between Russian cyber perspectives and practice. Russia's cyber posture, nested in Russia's concept of information warfare, is reflected in the offensive cyber operations launched by Russian government departments, whose institutional culture, expertise, and modus operandi have affected and will continue to affect Russia's cyber signature. This article reviews a combination of Russian primary and secondary open sources, scholarship of international researchers, and information available through online and traditional media. This article is further informed by an examination of modern publications, historical accounts, and unique, previously unpublished sources.

2. RUSSIA'S DOCTRINE AND STRATEGY ON CYBER SECURITY

A. A Shift in Russia's Understanding of Warfare

Over the past two decades, Russia's military and political leadership has undergone a fundamental modification of its conception of warfare and the role of cyber operations in this evolving view. Various scholars, such as Timothy Thomas, Martti J. Kari, Keir Giles, Oscar Jonsson, Brandon Valeriano, Benjamin Jensen, Ryan Maness, Stephen Blank, and Katri Pynnöniemi, have published seminal works in which they have analyzed various nuances of these dynamics (Thomas 2019; Kari 2019; Giles 2016;

Jonsson 2019; Jensen, Valeriano, and Maness 2019; Blank 2017; Kari and Pynnöniemi 2019; Medvedev 2015).[1] This section expounds this literature and serves as a reference guide to understand the trajectory of Russian cyber doctrine, cyber literature, and the assumptions that underpin them. It lays the foundation for the subsequent analysis on the evolution of Russia's cyber forces, which highlights the parallels between the existing doctrine and the Russian military scientific literature on one hand, and the organizational culture of Russia's main cyber departments and the nature of Russia's cyber operations on the other.

Russia's conceptualization of warfare has shifted from a general consensus that the baseline of warfare is armed violence to an agreement that the baseline for warfare has broadened to include a tailored amalgamation of armed violence and non-military measures (Chekinov and Bogdanov 2015a, 34; Chekinov and Bogdanov 2015b, 43; Jonsson 2019, 3–5; Gerasimov 2013; Burenok 2018, 61–66). Understanding these evolving nuances of Russia's military outlook is critical to Western decision-makers because the variation in the thinking of warfare between Moscow and the West also entails differences in understanding foreign policy signals and levers. Such differences may have wide-ranging consequences for deterring Russia and understanding Russia's red lines, and for facilitating the creation of a long-term strategy that addresses the causes of Russia's behavior.

Some of the terms that Western and Russian scholars have used to describe Moscow's shifting character of warfare include 'hybrid warfare', 'new generation warfare', 'the Gerasimov Doctrine', 'political warfare', 'hostile measures', 'cross-domain coercion', and 'gray zone tactics' (Chivvis 2017; Adamsky 2015; Morris et al. 2019; Galeotti 2018; Kofman 2016). Although these terms contain certain subtle and useful differences, they essentially attempt to capture an established understanding in Russia's strategic perceptions that warfare now includes non-military measures that an adversary can effectively use before, or in place of, overt military force (Jonsson 2019, Chapter 1).

It is worth noting that discussions over the employment of non-military measures in Russian warfare are not a novel phenomenon; however, these discussions were not adopted by a critical mass of Russia's military establishment until recent years. Russian military scholars have been expounding on the utility of such measures since before the Communist Revolution. During Napoleon's ill-fated campaign in Russia, Tsarist troops and Cossacks widely distributed leaflets aimed at lowering the morale of a conventionally superior enemy, including messages attempting to fracture the multinational invading coalition (Academy of Sciences 1962). The early Red Army similarly saw the utility of psychological warfare in applying pressure to populations behind the front. As a manual on military intelligence published during

1 The authors would like to express their gratitude to Martti J. Kari for his prompt and insightful comments on some of the arguments outlined in this article.

the ‘War Scare’ of the late 1920s states, “Political sentiment of the population in an enemy’s rear plays a big role in an opponent’s successful activities; because of this it’s extremely important to generate sentiments among populations against the enemy and use them to organize people’s uprisings and partisan detachments in the enemy’s rear” (Shil’bakh and Sventsitskiy, 1927). Additionally, Evgeny Messner, a pre-Revolutionary leading thinker in Russia’s strategic thought who wrote about the value and advantages of non-military measures, wrote extensively about the dissolution of boundaries between war and peace and the use of information operations to affect societal cohesion, which are reflected in the writings of a number of influential Russian military scholars who have outlined their vision of the evolving character of warfare since the 1990s (Jonsson 2019, 38–40; Gerasimov 2019; Chekinov and Bogdanov 2013). Despite the difference in means, as exemplified by the use of digital technologies today, the strategy undergirding modern Russian military cyberattacks and information operations was laid over a century earlier.

Despite the increasing number of articles on the use of non-military measures throughout the 1990s and 2000s, Russian military elites’ thinking changed most significantly between the early 2000s and the Ukraine crisis, when a consensus formed among senior Russian leaders and military theorists that the boundary between war and peace had become blurred and nonviolent measures of warfare could be so effective as to be considered violent, rendering them a tool of warfare (Jonsson 2019, 6–7, 153). The chief of Russia’s Armed Forces, Valery Gerasimov, wrote that the rules of warfare were changing and revolts modeled on the Arab Spring possibly presaged future wars where the protest potential of the non-military actors and the use of political, economic, and other non-military measures would be widely employed (Gerasimov 2014, 2013). Military scholars such as Colonel Chekinov and Lieutenant General Bogdanov further expounded on this argument, stating that the aggressive side will first use non-military measures, such as information technology aimed at engaging public institutions in a targeted country, including the media, cultural institutions, religious organizations, NGOs, and foreign-sponsored movements (Chekinov and Bogdanov 2013, 17). General Gerasimov reemphasized the employment of mixed tactics and the maintenance of asymmetrical and classic potential at the 2019 conference of the Russian Academy of Military Sciences. He noted the changing character of war and the evolving “coordinated use of military and non-military measures” and even suggested the primacy of non-military measures over military power, used only when impossible “to achieve the goals set by non-military methods” (Gerasimov 2019).

Recent amendments of Russia’s main strategic documents also reflect an evolving view of warfare. The 2010 Russian Military Doctrine stated that integrated non-military and military means is a characteristic of modern military conflicts (President of Russia 2010). The updated 2014 doctrine reinforced this concept and listed it as the

first characteristic of modern military conflicts: "the integrated use of military force, political, economic, informational and other non-military measures implemented with widespread use of the protest potential of the population and special operations forces" (Rossiyskaya Gazeta 2014). The 2013 Foreign Policy Concept listed economic, scientific, and IT factors as being important as military capabilities to influence politics in a given state (Ministry of Foreign Affairs 2013). These speeches and doctrinal documents illustrate the conceptual flip that evolved in Russia's perceptions of modern warfare.

B. Russia's Official Views on Information Warfare

Outlining the contours of Russia's view on warfare is critical for grasping Russia's cyber strategy because Russia's view on cybersecurity is nested in Russia's evolving understanding of the nature of war and is shaped by its concept of information warfare.[2] Cybersecurity is perceived as a Western notion in Russian debates, while the semantic Russian equivalent is information security (*informatsionnaya bezopastnost*). Military scholars and official documents present slightly varying definitions of information warfare and information security, but it is generally well-established that information security is a component of information warfare, which is a term that has a technical as well as a psychological or cognitive component. Information warfare is an integral part of interstate conflict and its aim is to establish information superiority over the adversary by using technical and psychological means, while cyber operations are a mechanism used by the state to dominate the information environment, which is considered a domain of warfare (Thomas 2019, 5–5, 7–8, 7–9; Connell and Vogler 2017, 3). Russia's Ministry of Defense 2011 Concept on the Activities of the Armed Forces of the Russian Federation in the Information Space provided a clear definition of information warfare:

> …the confrontation between two or more states in the information space with the purpose of inflicting damage to information systems, processes and resources, critical and other structures, undermining the political, economic and social systems, a massive psychological manipulation of the population to destabilize the state and society, as well as coercing the state to take decisions for the benefit of the opposing force (Ministry of Defense of the Russian Federation 2011).

This definition emphasizes the two main elements of information warfare, namely the technical element of information infrastructure, which consists of a mix of "technical tools and systems of formation, creation, transformation, transmission, usage and storage of information" (roughly corresponding to issues pertaining to information

2 Russia's military literature and doctrine use three terms that can be roughly translated as information warfare. These are *informatsionnoe protivoborstvo* (information struggle or information confrontation), *informatsionnaya voina* (information war) and *informatsionnaya borba* (information fight). Explaining the nuances of each term is beyond the scope of this paper and for the purposes of this research, we will use the translation "information warfare". Also see Giles 2016, p. 7, footnote 8.

and cybersecurity in the West), and the psychological component of information warfare, which involves cognitively influencing the population and decision-makers of the opposing state to erode their will to fight and their decision-making structures and processes (Ministry of Defense of the Russian Federation 2011; Chekinov and Bogdanov 2015b, 45).

The information sphere and the concept of information warfare fits well within Russia's understanding of the changing character of war because, as General Gerasimov asserted, "without having clearly defined national borders, [the information sphere] provides the possibility of remote, covert influence not only on critical information infrastructures, but also on the country's population, directly affecting the state's national security." These characteristics render studying issues of preparation and conduct of informational activities "the most important task of military science" (Gerasimov 2019). Considering its multifaceted and unconventional nature, information warfare, and by extension cyber operations, may commence prior to the official announcement of war and can be deployed to achieve political objectives without resorting to the use of military force (President of Russia 2010).

C. Main Threats Posed in the Information Sphere

The threat posed by information means has gradually gained prominence in Russian doctrine since the start of the 21st century. In line with the Soviet tradition of portraying Russia as a besieged fortress defending itself against constant internal and external threats, Moscow also views the struggle in the information sphere as constant and unending (Kari 2019, 84, 72–6; Kari and Pynnöniemi 2019, 21; Connell and Vogler 2017). The 2000 National Security Concept highlighted that Russia's national security is threatened in the information sphere by countries that are attempting to dominate the information sphere while developing their concept of information wars. The Security Concept presented a holistic understanding of the term by focusing on threats that are related to both the technical and the psychological aspects of information warfare (Ministry of Foreign Affairs of the Russian Federation 2000). Russia's 2010 Military Doctrine further elevated the status of information warfare and signaled a shift in the formal understanding of threats to the nation by listing the increasing role of information warfare for the first time as a characteristic of contemporary military conflicts and the imperative for Russia's military to develop forces and means of information warfare (President of Russia 2010).

The 2000 and the 2016 Russian Information Security Doctrines further codified Russia's official view on the role of information threats in contemporary warfare (Table 1). The 2000 doctrine provided a broad definition of the information sphere, which is a "combination of information, information infrastructure, entities involved in the collection, generation, distribution and use of information, as well as a system

for regulating the resulting public relations" (*Nezavisimaya Gazeta* 2000; President of Russia 2016). This definition is in line with the understanding that Russia's information sphere includes a technical and a cognitive component. Based on this broad definition, the concept includes a wide array of threats to information security. They range from more technical threats, such as threats to the security of information and telecommunication facilities and systems that include "the introduction of electronic devices for intercepting information in the technical means of processing, storing and transmitting information," and broader threats to societal cohesion, such as "decrease in the spiritual, moral and creative potential of the Russian population" (*Nezavisimaya Gazeta* 2000).

The 2013 Security Council's Basic Principles on International Information Security confirmed this broad understanding and the panoply of threats related to information security and saw information technology as a weapon that can be used for political and military purposes to violate a state's sovereignty and territorial integrity (Security Council of the Russian Federation 2013). The updated 2016 Information Security Doctrine continued in the spirit of its conceptual predecessors by reemphasizing the growing threat posed to Russia in the information sphere by various adversaries (President of Russia 2016). The doctrine emphasized increasing threats emanating from the information cognitive space, primarily driven by foreign actors, and their effects on social values and stability (President of Russia 2016). These documents illustrate the belief that Russia's posture in the information sphere is shaped in response to threats to Russia that are forcing the state into defending itself.

D. Russia's Doctrinal Response to Threats in the Information Sphere: Defensive and Cooperative Posture

Russia's officially expressed strategy to manage threats in the information sphere is as multifaceted and broad as the threats themselves, yet the strategy is generally consistent in its omission of offensive or adversarial actions (Table I). In official documents, the government lists policy goals that outline a primarily defensive and collaborative posture designed in response to aggressive adversaries and entities that threaten Russia, which aims to contain or prevent aggression in cyberspace through legal frameworks and partners. Such national-level policies include the "development and adoption of regulatory legal acts of the Russian Federation establishing the liability of legal entities and individuals for unauthorized access to information, its illegal copying, distortion and illegal use" and enhancement of "the security of critical information infrastructure" (*Nezavisimaya Gazeta* 2000; President of Russia 2016). International policy recommendations range from the "formation of a system of international information security" to "the formation of mechanisms for international cooperation in countering the threats of the use of information and communication technologies for terrorist purposes" (Security Council of the Russian Federation 2013).

TABLE I. A SELECTED LIST OF MAIN THREATS AND RECOMMENDED POLICY RESPONSES AS OUTLINED IN MAIN RUSSIAN INFORMATION SECURITY DOCUMENTS

Document	Threats		Recommended Policy Response
	Psychological	Technical	
Information Security Doctrine (Nezavisimaya Gazeta 2000)	• irrational, excessive restriction of access to socially necessary information; unlawful use of special means of influence • ousting Russian news agencies, the media from the domestic information market and increasing the dependence of the spiritual, economic and political spheres of public life in Russia on foreign information structures • a decrease in the spiritual, moral and creative potential of the Russian population	• development and distribution of programs that interfere with the normal functioning of information and information and telecommunication systems, including information protection systems • compromise of keys and means of cryptographic information protection • destruction, damage, or theft of machines and other storage media	• introduction of amendments and addenda to the legislation of the Russian Federation regulating relations in the field of ensuring information security in order to create and improve the system of ensuring information security of the Russian Federation • clarification of the status of foreign news agencies, media and journalists, as well as investors when attracting foreigners' investments for the development of information infrastructure in Russia; • legislative priority for the development of national communications networks and domestic production of space communications satellites
Conceptual Views on the Activities of the Armed Forces in the Information Space (Ministry of Defense 2011)	• threats of a political nature in the information space	• widespread use of computer technology in command and control systems of troops and weapons	The activities of the Armed Forces of the Russian Federation in the information space are built on the basis of a set of principles: legality, cooperation with friendly states and international organizations; and containment and prevention of military conflicts in the information space
Convention on Ensuring International Information Security (Ministry of Foreign Affairs 2011)	• factors creating a danger to the individual, society, state and their interests in the information space • actions in the information space in order to undermine the political, economic and social systems of another state, psychological treatment of the population, destabilizing society • using the information infrastructure to disseminate information that incites ethnic, racial and inter-confessional enmity, racist and xenophobic written materials	• targeted destructive impact in the information space on the critical structures of another state • countering access to the latest information and communication technologies, creating conditions for technological dependence in the field of informatization to the detriment of other states • information expansion, acquisition of control over the national information resources of another state	State parties should: • maintain international peace and security and promote international economic stability and progress, the general welfare of peoples and international cooperation, free from discrimination • refrain from developing and adopting plans and doctrines that can provoke an increase in threats in the information space, as well as cause tensions between states and the emergence of "information wars" • refrain from any action aimed at the complete or partial violation of the integrity of the information space of another state
Basic Principles for State Policy in the Field of International Information Security until 2020 (Security Council 2013)	• carrying out hostile acts and acts of aggression aimed at discrediting sovereignty, violating the territorial integrity of states and posing a threat to international peace, security and strategic stability • interfering in the internal affairs of sovereign states, disturbing public order, inciting interethnic hostility	• destroy elements of critical information infrastructure • crimes, including those related to unlawful access to computer information, with the creation, use and distribution of malicious computer programs	• formation of a system of international information security at the bilateral, multilateral, regional and global levels • creating conditions to reduce the risk of using information and communication technologies for hostile acts and acts of aggression aimed at discrediting sovereignty, violating the territorial integrity of states and posing a threat to international peace, security and strategic stability
Information Security Doctrine (President of Russia 2016)	• increasing use by the special services of individual states of information and psychological influence aimed at destabilizing the domestic political and social situation in various regions of the world and leading to the undermining of sovereignty and territorial integrity • increase in materials in foreign media containing a biased assessment of the government policy of the Russian Federation	• increase in the scale and coordination of computer attacks on objects of critical information infrastructure, increased intelligence activities of foreign states against the Russian Federation, as well as an increase in threats to the use of information technologies in order to cause damage territorial sovereignty integrity, political and social stability of the Russian Federation	• strategic deterrence and prevention of military conflicts that may arise as a result of the use of information technology; forecasting, detection and assessment of information threats, including threats to the Armed Forces of the Russian Federation in the information sphere • neutralization of information-psychological impact, including aimed at undermining the historical foundations and patriotic traditions associated with the defense of the Fatherland

E. Cybersecurity beyond Russia's Doctrine: The Value of Cyber Weapons

Although Russia does not have an explicit cybersecurity doctrine and its formal documents discussing Russia's posture in the information sphere show a primarily defensive posture, Russia's theoretical military literature provides additional useful insights into the role of cyber capabilities, especially offensive cyber capabilities, in Russia's view of conflict. Military scholars elaborate on the appositeness of cyber weapons in modern warfare, on their versatility and effectiveness, and on their affordability. Offensive cyber capabilities fit within the concept of information warfare because cyberspace allows for blurring of the boundaries between war and peace, as damage can be inflicted on an adversary during peace time without crossing the threshold of armed conflict or declaring war as a legal act. Enabled by a lack of clear legal framework to serve as the foundation for prosecuting the perpetrators of cyber operations, an adversary can conduct hostile or destructive cyber operations from any location and can weaken the enemy's ability to defend themselves and retaliate (Vorob'ev and Kiselev 2013, 33–4; Kuznetsov et al. 2018, Parshin and Bashkirov 2019, 5; Antonovich 2011; Thomas 2010, 287; Starodubtsev, Bukharin and Semyonov 2012; Jonsson 2019, 108). Another military virtue of cyber weapons, as then First Deputy Chief of the General Staff, General Aleksander Burutin, and others argued, is that these weapons can help an adversary achieve information supremacy without crossing borders or establishing physical presence on the enemy's territory (Thomas 2010, 287; Parshin and Bashkirov 2019, 6). Even perhaps more importantly for Russia, offensive cyber capabilities can be considered as asymmetric actions that can help a technologically and economically weaker state (which Russia considers itself to be vis-à-vis the United States) to neutralize a stronger opponent (Selivanov 2020, 50; Kari 2019; Burenok 2018). Offensive actions in cyberspace may also be preferable to defensive ones, as the former are deemed faster than the latter (Mikryunov 2015, 117).

Russian military scientists have repeatedly noted the destructive capacity and versatility of cyber weapons, which can be employed against civilian, military, and government targets. In line with Russia's doctrinal understanding of information warfare, scholars argue that the deployment of cyber weapons can affect adversaries' infrastructure as well as their psychology. In an article prepared on behalf of the Defense Ministry, Bazylev et al. elaborated on the technical impact of cyber weapons and argued that such weapons can critically affect facilities in the transportation or energy sectors, and can even lead to a financial crisis (Bazylev et al. 2012, 24–25, Jonsson 2019, 108). Military scientists Kiselev and Kostenko expounded that cyber weapons can endanger not only critical infrastructure elements such as supervisory control and data acquisition (SCADA) systems and smart power systems but also military systems (Kiselev and Kostenko 2015, 4). During conflict, such weapons can render the enemy's control infrastructure dysfunctional and the higher the level of automation of objects and processes of the targets, the greater results that can be

achieved because of the existence of vulnerabilities in these systems (Starodubtsev, Bukharin and Semyonov 2012, 29-30; Kuznetsov et al. 2018, 5). In addition to their technological effects, these weapons can "completely disorganize state and military administration, demoralize and disorient the population, and create mass panic" (Bazylev et al. 2012, 24-5, Jonsson 2019, 108). Former Deputy chief of the General Staff, Colonel-General Anatoliy Nogovitsyn, and others further elaborated on the offensive role of cyber tools and their dual impact, explaining that they can destroy military, administrative, and industrial sites, while also inflicting information and psychological damage on the enemy's troops, leadership, and population (Thomas 2010, 287; Parshin and Bashkirov 2019, 4, 8-9).

Another positive characteristic of cyber weapons discussed by military scientists is their relatively low cost. The development and creation of such weapons is estimated to be much cheaper than other types of weapons, while the use of either leads to comparable damage (Parshin and Bashkirov 2019, 6; Romashkina and Kildobskiy 2015, 134; Putin 2012; Jonsson 2019, 109). A study further elaborates that the total defeat of the information infrastructures of major powers such as the United States or Russia could be conducted by up to 600 "information warriors." Training these warriors and executing the actual attack would take about two years and cost no more than 100 million dollars (Bazylev et al. 2012, 24-5). Another potential reason for the relative affordability of such weapons is that operational plans for their use may be developed by non-military experts (Starodubtsev, Bukharin, and Semyonov 2012). Despite the lack of explicit discussion on specific Russian cyber operations or developments of cyber weapons, the literature offers certain clues as to how Russia's military elite views cyber warfare and offensive cyber capabilities on a theoretical level, which demonstrates a realization of the value of cyber weapons as having high levels of effectiveness and versatility, high affordability, and fitting within the current character of warfare.

The analysis of Russia's doctrine, speeches of Russia's elite, and the military scientific literature paints a general picture of Russia's vision of cybersecurity, which is situated in Russia's understanding of information security and information warfare. Although Russia's official documents describe Russia's view on information warfare as defensive, Russia's military literature shows an active debate on the value of developing and fielding both defensive and offensive cyber capabilities. The interest in discussing cyber weapons in Russian military journals, coupled with proactive Western cyber policies, such as the strategy of persistent engagement and the concept of defending forward that is endorsed by U.S. Cyber Command, may provide sufficient justification that will prompt the Russian leadership to formally include the development and deployment of cyber weapons in its information warfare doctrine (U.S. Cyber Command 2018). On the other hand, the continuous omission of an

official endorsement of offensive cyber capabilities in its doctrine allows the Russian government to claim plausible deniability and maintain a narrative (as questionable as that narrative is among Western observers) of a defensive power under threat by an aggressive West – a classic justification for a number of Russian policies, including investments in military modernization.

To further understand Russia's cyber strategy and policy, this article will examine the evolution and institutional character of the structures of Russia's government that are involved in the conduct of Russia's information and cyber operations, which appear to follow Russian doctrine and literature on the importance of developing cyber capabilities that have both technical and psychological effects.

3. THE EVOLUTION OF FSB AND GRU CYBER AND INFORMATION OPERATIONS

A. The Initial Years of Russia's Cyber Operations: The FSB and Non-state Actors

Throughout most of post-Soviet Russia, the Federal Security Service (FSB) maintained the "commanding heights" of external cyber operations. In the unregulated space of the Russian internet in the 1990s and early 2000s, the FSB developed relationships that helped it coopt or coerce independent Russian hackers and specialists into cyber operations. Layers of unofficial hackers helped circumvent the human capital challenges that long impaired Russia's early development of cyber-capable cadres. For instance, an anonymous source within one of the FSB's leading hacking departments, the Center for Information Security (CIS), claimed that the unit employed illegal hackers to make up for its staffing deficiencies (Turovsky 2018, 149), while another source claimed that one of the leading CIS hackers, when recruiting external support, often created an "atmosphere that Russia needed help," even more so after the 1990s, when attacks against banks in Europe and the U.S. could help alleviate financial shortfalls (Turovsky and Rothrock, 2018). The FSB's inheritance of the bulk of the Federal Agency of Government Communications and Information (FAPSI), a loose analog to the U.S. National Security Agency that was disbanded in 2003, alongside the Kvant Scientific Research Institute that has assisted the FSB's technological research for over a decade, provided the FSB with a significant advantage in fostering an offensive cyber capability (U.S. Department of The Treasury 2018). As longtime cybersecurity correspondent Andy Greenberg wrote of the period, "…the GRU [the Main Intelligence Directorate of Russia's military] had taken a backseat to the FSB throughout Russia's inchoate cyberwars in Estonia and Georgia, relegated to traditional intelligence in direct support of the military rather than the exciting new realm of digital offensive operations" (Greenberg 2019, 236).

For a while, this fluid basis for cyber operations served Moscow's interests. The "Siberian Network Brigade," a group of Russian students from Tomsk University, enjoyed legal cover from their local FSB branch as they launched Distributed Denial of Service (DDoS) attacks against Chechen websites in the early 2000s (Gazeta.ru 2006; Newsru.com 2002). The renowned example of attacks against Estonia in 2007 similarly involved an amorphous coalition of state-sponsored hacking that mostly continues to defy firm attribution. At the same time, malware most likely associated with the FSB penetrated U.S. defense networks to facilitate one of the most significant breaches of classified data in history (Council on Foreign Relations 2008). Throughout the early 2000s, there was little reason for Moscow to seriously consider an alternative to an FSB-led cyber program, and the latter's prominence in executive leadership circles ensured its lead. As Keir Giles noted in 2011, the prospect of "information troops" in Russia's military, which would include cyber operations, was officially discounted by the FSB at the time (Giles 2011).

Ironically, some of the FSB's earlier operations perhaps helped bring about the eventual ascension of the Russian military's cyber program, which languished under post-Soviet malaise, meager budgets, and personnel deficiencies. The cyberattacks on Estonia and Georgia, plus the exploitation of U.S. defense networks by Russia and other states, prompted the U.S. to strengthen its own military program, most notably with the foundation of the U.S. Cyber Command in 2009. Other events concurrent to the Cyber Command's development, such as the revelations surrounding the unprecedentedly sophisticated "Stuxnet" malware targeting Iran's nuclear program, reinvigorated concerns among Russian security and defense observers about U.S. predominance in cyberspace. U.S. efforts to apparently militarize its growing cyber capabilities necessitated that Moscow redouble efforts to improve those within its military. Unproductive negotiations between Russian and Western interlocutors about regulating evolving cyber capabilities, caught in fundamental divides on issues like international internet governance, dwindled the prospect of "cyber arms control" between Moscow and its perceived adversaries (Krikunov 2011, 32–7; Tikk and Kerttunen 2018; Kavanaugh 2015). While loose, ad-hoc coalitions of cyber actors outside the state's direct purview may have been sufficient for Russia's earlier cyber ambitions, the apparently widening gap in capabilities between it and other states and alliances, chiefly NATO, exacerbated preexisting fears about unpreparedness for what was increasingly viewed as an inevitable information confrontation with the West.

B. The Advent of the GRU to Information Warfare

In mid-2013, after receiving presidential approval, Russian Defense Minister Sergey Shoygu launched a "big hunt" for programmers to fill the ranks of new "military science units" (*voennye nauchnye roty*) that would advance the military's research and development through the coming years, with an emphasis on cyber operations,

signals intelligence, and electronic warfare.[3] Of the four original science companies, one belonged to the GRU, which had an unmistakable focus on computing and information technology.[4] In May the following year, sources within Russia's Ministry of Defense announced the establishment of an "information operations force" (*voyska informatsionnykh operatsiy*), which, according to the Russian press, was partly predicated on the growth brought through the science units and the development of which was catalyzed by the leaks of classified U.S. programs by Edward Snowden (TASS 2014; Saltykov 2014). Moreover, the 2014 Military Doctrine listed the "development of forces and means of information confrontation" as a main task of equipping Russia's modernizing armed forces (*Rossiyskaya Gazet*a 2014). By early 2017, Shoygu was confident enough in the force to announce its readiness before Russia's national legislature. Between his "big hunt" and 2017, the attribution of Russia's most significant cyber operations to the GRU by Western intelligence agencies and a range of private cybersecurity and investigative organizations evidenced the arrival of the GRU as the probable leader in large-scale cyberattacks.

As the Main Intelligence Directorate of Russia's General Staff came to the fore in offensive cyber operations, it brought with it a culture of aggression and recklessness; the same day that the GRU's Main Center for Special Technologies launched the costliest cyberattack to date, the 'NotPetya' wiperware that led to over $10 billion in damages, a car bomb in Ukraine's capital killed a Ukrainian special forces officer (Greenberg 2017; Nakashima 2018).

The GRU's seemingly high tolerance for operational risk is in many ways incongruent with the traditionally furtive realm of cyber operations, which consist far more often of quiet espionage efforts than large-scale attacks. A former FSB cyber officer who was arrested in late 2016, possibly in an effort to expose GRU hackers by leaking information about them, claimed that the GRU "impertinently, roughly, and brutishly breaks into servers," which always led to their attribution (Turovsky 2018, 198). Whatever the GRU's apparent missteps, the organization at least publicly maintains President Putin's confidence, and the continuous attribution of Russian cyber and

3 For example, *Rossiyskaya Gazeta*, a state-controlled press outlet, ran an article in 2013 titled "Private [military rank] Hacker" (*ryadovoy khaker*) that accompanied the rollout of the science units (Gavrilov 2013). Moreover, as journalists with Meduza acutely noticed, science-unit recruitment was likely bolstered by a 40-part TV show aired by the Zvezda network that glamorized new recruits' work in a Russian military cyber unit (Turovsky 2016). Though most science units conduct some research outside of computer science or information technology, almost all have some cyber research component, which is certainly true of the four original units established in 2013. Moreover, the newest such units, assigned to the 'ERA' technopolis based in Anapa, Russia, concentrate on cyber-relevant projects, judging from official documents and press reporting (Ministry of Defense of the Russian Federation 2018; Ren.tv 2019).

4 For example, the GRU's science unit maintained a stand at the military's 2015 "Innovation Day," where it displayed materials with a clear focus on computer science research (Livejournal.com 2015). Additionally, an archived copy of an anonymous resume from a former member of that unit demonstrates an exclusive background in computer programming. According to the official website of Bauman State Technical University, the GRU's science company is based in Zagoryanskiy and is designated as Unit Number 36360 (Bauman Moscow State Technical University).

information operations to it show that the GRU is likely to continue conducting these campaigns (Balforth 2018). The graduates of computer science programs brought into the GRU's ranks through its own science unit(s) and other initiatives are most likely distant from their counterparts in Russian "*spetsnaz*" units. As Andrey Soldatov explained, the "stereotypical portraiture of a GRU hacker" is "far from universal," as the organization recruits non-military types "conscripted for their services with little choice in the matter" (Greenberg 2019, 242). But, to the extent science unit(s) recruiting advertisements, which feature a Kalashnikov assault rifle propped next to a computer, suggest the culture into which these recruits enter, GRU operators are likely to continue meshing a daring culture of special operations with digital activity, an undoubtedly alluring prospect for at least some of Russia's youth (*Nauchnaya Rota* REB 2015).[5] The importance that Russian defense officials place on their work only reinforces this aura of exigency and adventure. A vice-admiral who reportedly delivered a science-unit recruiting pitch to university students in 2013 compared their future work to the Soviet Union's development of an atomic bomb, which echoed a similar comparison by Moscow's foremost cyber-diplomat, Andrey Krutskikh, in 2016 (Habr.com 2013; Ignatius 2017).

C. GRU's Organizational Culture and the Conduct of Information Operations

Another aspect of GRU culture has driven its adoption of cyber operations and has largely been unexplored: its history and growing fixation on information operations. Contrary to most of the GRU's cyber units,[6] its information operations forces have a deep history; the Red Army dedicated a force to "special propaganda" (*spetsprop*) shortly before World War II, and these forces represent a component of Russian information warfare as indispensable as technical capabilities. *Spetsprop* units broadcasted messages and distributed leaflets and products to enemy forces to reduce their morale and entice surrender, and they worked to influence civilian populations behind the frontlines and when promoting civil-military operations in the wake of advancing armies, though efforts to foster public support were quickly undone by mass arrests and deportations. After 1991, these units were rebranded and placed exclusively under the GRU.[7] The GRU organized many of these specialists into eight "psychological operations groups" during the throes of the first Chechen War and dispersed them throughout Russia's military districts (Kozlov 2010, 176).

Nonetheless, disappointment in the military's ability to counter perceived Western information warfare aimed at Russia during the Georgian War (Iasiello 2017) drove

5 The same unit that posted the above recruiting video was tangentially associated with the GRU's "Fancy Bear" hacking team when, in 2015, it posted another recruiting video that featured the emblem associated with the group, though it was subsequently taken down (Turovskiy 2016).

6 The exception to this is Unit 26165, or the 85th Main Special Service Center, which was founded in the 1970s to conduct signals intelligence.

7 During the Soviet era, special propaganda units belonged to the Main Military Political Directorate (GlavPUr). Following the collapse of the Soviet Union, they were renamed "Centers for Foreign Military Information and Communication" (*Argumenty Vremeni* 2018).

defense officials to rejuvenate *spetsprop* in the 21st century. Officials realized that modern propaganda, like that seen to be used by NATO, needed to be digital. An official in the GRU's information operations training pipeline,[8] for instance, claimed in accordance with the Russian information warfare doctrine sometime after the Georgian War:

> The features of modern information confrontation show that it is [as] directed at both information-technical systems … as it is on human psychology. Activity against an enemy is organized and conducted in two aspects (directions): technological and psychological (Cheshuin 2009).

New aspects of information warfare, such as DDoS attacks, would be introduced to the information operations faculty of Russia's Military University of the Defense Ministry following the Georgian War and combined with old operational practices, such as disinformation (Cheshuin 2009).

As much as cyberattacks provided a new means for asymmetric tactics, modern information communications technology also provided the GRU with an updated arena for propaganda techniques that extended back to the foundation of *spetsprop*. Roughly 80 years before GRU specialists attempted to stir Polish-Ukrainian tensions in Lviv through social media, Red Army propagandists pitted the two nationalities against one another in the same region to ease the Soviet invasion of eastern Poland at the onset of World War II (Diresta and Grossman 2019, 55; Repko 1999, 267). Similar to special propagandists' use of German radio networks to entice surrender during that war, the modern GRU orchestrated the demoralizing text messages that have been sent to Ukrainian soldiers since 2014 (Burtsev 1981, 166–67; Tribun 2018).

These units' activity since the early 2000s demonstrates their "digitalization," including their eventual involvement in cyberattacks. During the Second Chechen War, they launched an unsophisticated "e-newspaper" titled "Morning" (*Utro*) to color events surrounding the conflict (Kompromat.ru 2002). The GRU's efforts to conduct online influence operations probably evolved somewhat by the start of the Ukraine crisis in 2014, though their use of a Facebook primer containing basic instructions on using the platform indicates operators were still somewhat unfamiliar with waging an internet-based information war (Nakashima 2017). Only a year later, however, the GRU combined cyberattacks, primarily against France's TV5 Le Monde, with influence operations through ISIS social media cutouts as part of its "CyberCaliphate" campaign (Sengupta 2018). Like the apparent recklessness in hacking used to support the campaign, CyberCaliphate involved direct physical threats via social media

8 The "Faculty of Foreign Military Information" at Russia's "Military University of the Defense Ministry" (VUMO) has long served as the main training pipeline for Soviet and Russian psychological warfare units, and its history extends back to the foundation of special propaganda. According to information on a Russian website on academic institutions in the Moscow area, the faculty directly sends its graduates to the GRU (Moscow-Russia.ru).

against U.S. military spouses, exemplifying that digital aggression would carry over into influence operations (Slatter 2018). The involvement of the nucleus of the GRU's psychological warfare apparatus, the 72nd Special Service Center (Unit 54777), demonstrated that information operation specialists would work alongside GRU cyber units throughout the campaign (Troianovski and Nakashima 2018).

According to Western intelligence officials, the 72nd Special Service Center (Unit 54777) has been in lock-step with GRU hackers since at least 2014, complementing cyberattacks with digital information operations through proxies and front organizations (Troianovski and Nakashima 2018). Before the Ukraine crisis, Unit 54777 had 80 specialists split among five sections: a Center for Foreign Military Information; a department for organizing and conducting psychological or information operations; a department for organizing "teleradio" broadcasts; a department for working with mass media; and an editorial-publications department.[9] The unit sent advisors to Russia's various military branches, such as the ground forces and navy, and levels of command that reached from GRU leadership to tactical units manning frontline loudspeaker vehicles.[10] This plausibly served as a prototype for the "information confrontation" chain-of-command revealed by Gerasimov during a staff exercise in 2016 (Izvestiya 2016). Though unverified, Ukrainian accounts of regional GRU information operations units conducting cyber and electronic warfare operations probably demonstrate the capabilities of local commands to conduct operations at lower echelons (Tribun 2018).

D. GRU's Organizational Culture and the Conduct of Technical Cyber Operations

While the GRU's cyberattacks have attracted much research and analysis throughout the past six years, less effort has been given to discerning how the organization's history influences contemporary operations. Russian military intelligence's cyber operations are rooted in the history of its technical intelligence that, while perhaps not as extensive as that of information operations, predates World War I. Technical intelligence, primarily cryptography and signals intelligence, underwent its most significant and expansive development during the Soviet period. Early Soviet military leadership recognized its importance, expanding the number of "radio-reconnaissance stations" throughout the U.S.S.R. and abroad throughout the 1920s, allowing signals intelligence to play a central role in the Sino-Soviet conflict in 1929 (Kozlov 2013, 411). Soviet military signals intelligence and cryptography achieved notable prewar successes in the Far East, surpassing British and equaling U.S. collection capabilities in that theater by 1939 (Haslam 2015, 98). Despite at least occasional effectiveness, the Soviet military's early technical intelligence capabilities mostly existed in the shadow of the internal security services, such as the subordination of decryption specialists to the Joint State Political Directorate (OGPU) (Larin 2017, 65). World

9 Discussion with experts, May 2018. Helsinki.
10 Ibid.

War II prompted breakneck growth to Soviet military technical intelligence, and – by 1942 – military cryptologists successfully cracked the German military's "Enigma" machine, and eventually began intercepting and deciphering German communications with enough regularity to force German signal officers to forbid marking "the Fuhrer's radio messages in any special way" (Kahn 1996, 649). Throughout ebbs and flows in terms of political influence, resources, and relations with the more powerful KGB, the GRU continued to expand its signals intelligence capabilities during the Cold War; by the Gorbachev era, the Soviet military possessed 40 signals intelligence regiments, 170 battalions, and over 700 companies (Andrew and Mitrokhin 1999, 353).

One of the most significant developments for Soviet military signals intelligence during the late Cold War was the establishment of the 85th Main Special Service Center (Unit 26165), which was responsible for GRU cryptography through a variety of technical means, including the "Bulat" computer system (Shevyakin 2014, 104). The center's independence from the GRU's signals intelligence directorate and direct subordination to GRU leadership exemplified the importance of their work. Whatever the center's prominence in the Cold War, it very likely suffered from the same post-Soviet reductions that affected the broader Russian military and its intelligence capabilities. Nonetheless, officers like Viktor Netyshko, who would eventually head the center during its efforts to influence the 2016 U.S. presidential election, ensured that the 85th would continue its mission and development of cyber capabilities no matter the shortfalls, albeit at a reduced capacity. Fewer resources, including access to recruits during a period when the military was supposed to drastically expand its cyber specialists, likely influenced the eventual agreement between Netyshko and the FSB in 2017 to jointly prepare recruits at the latter's cryptography institute probably in part for entry into the military's science unit(s) (Moscow State Budgetary General Education Institute 2017).[11] In the meantime, future leaders of the center pursued scientific and academic research related to the kind of computer science needed to advance cyber operations. In 2003, Netyshko defended a dissertation related to the academic specialty "Mathematical and Programming Software of Computers, Complexes, and Computer Networks," and in 2010 he served as an opponent for a dissertation on computer hacking (Turovsky 2018, 195). Sergey Gizunov, who preceded Netyshko as the center's commander and who simultaneously taught computer science, was awarded the title "Laureate of the Government of the Russian Federation in the Field of Science and Technology" in 2008 (*Rossiyskaya Gazeta* 2009). Gizunov's promotion to GRU deputy director in 2015 likely evidences the growing influence of technically proficient officers experienced in cyber operations.

The 85th Special Service Center, however, represents only a part of the GRU's offensive cyber apparatus. The Main Center for Special Technologies (Unit 74455) has similarly captured significant attention surrounding its involvement in the effort

11 As described in the document, the FSB's Institute of Cryptography, Communications, and Informatics Academy would prepare recruits for entry into the FSB's academy and "targeted groups of military units at technical universities" (*tselevye gruppy Voyskovoy chasti VUZov tekhnicheskovo profilya*).

to influence the 2016 U.S. presidential election and the "NotPetya" cyberattack the following year. Unit 74455's historical roots are far shallower than Unit 26165's history as part of Soviet signals intelligence, and the former's establishment probably reflected the mounting importance of strictly computer-based operations to Russia's military leadership. Its officers are seemingly also closely connected to military computer science research; a commander of one of Unit 74455's departments reportedly teaches "applied information technology" at the Mozhayskiy Military-Space Academy (Faizova et al. 2018). An apparent link between Unit 74455 and the 4th Central Scientific Research Institute, a defense ministry entity historically dedicated to the strategic missile forces, potentially couples GRU hackers with research relevant to evolving military theory and strategy surrounding cyber operations.[12] The continued authorship of articles between 2008 and 2018 related to cyber capabilities in a journal titled *Information Wars* by 4th Central Scientific Research Institute officials probably indicates a growing interest by the organization in cyber issues, such as a 2018 article titled "Threat Models of Joint Information-Technical and Information-Psychological Effects in Hybrid Wars" (Antonov et al. 2018). At the same time, operations attributed to Unit 74455 against Ukrainian, European, and Western targets demonstrated an increasing sophistication that likely partly stemmed from better resourcing and staffing. Marina Kotofil, an industrial control systems expert, remarked about the difference between the 2015 and 2016 operations to disrupt Ukrainian energy grids, "In 2015, they were like a group of brutal street fighters … in 2016, they were ninjas" (Greenberg 2019, 133).

E. Implications of the Rise of Russian Military Cyber and Information Operations for Future State-Sponsored Activity

The fall 2019 cyberattacks committed by the GRU against Georgia exhibited the inseparability of the technical from information elements of contemporary information warfare, using sophisticated malware to black out television and websites while disseminating an image of Georgia's former president, who was indicted on corruption charges in 2013, claiming he would return (Greenberg 2020). This integration is very likely to continue in future campaigns, such as potential cyber flashpoints between Russia and the West surrounding upcoming presidential and parliamentary elections in 2020, and deepening political and societal divisions within several of those states to provide Russian state-sponsored actors with an opportunity to continue undermining perceived adversaries through digital means. As these vulnerabilities to cyber and information operations have worsened, Moscow has likely continued to hone and expand the cyber capabilities to exploit them. A late 2019 report by Check Point Software Technologies, for instance, claimed that state-sponsored actors invested a "significant amount of money and effort" in the first half of 2019 to develop "large-

12 One of the servers used by Unit 74455 to conduct operations related to the effort to undermine the 2016 U.S. presidential elections was based at the same address as the 4th Central Scientific Research Institute (Kritukov 2018). Moreover, a document related to a military court decision in 2010 revealed the transfer of an employee of the institute probably to Unit 74455 to lead "department 24" (Znamensk Garrison Military Court 2011).

scale espionage capabilities," which the firm concluded was an unprecedented investment by Russia in "offensive cyberspace" (Doffman 2019). The imperative to understand these capabilities has perhaps never been greater, and studying the organizational culture and history of the actors responsible for carrying out cyber and information operations offers unparalleled insight into the motivation, strategy, and methods guiding their respective efforts.

Given the consequences and reach of the GRU's cyber and information operations, which range from debilitating a swath of global shipping through wiperware to attempting to stoke racial tensions in the U.S., understanding the actors behind this activity on a more specific level is critical for anticipating potential future efforts and understanding how to address them (Greenberg 2019, 174–89; Digital Forensics Research Lab 2018). In part, this involves historical research on Russian intelligence. While countless Western publications continue to discuss the Gerasimov Doctrine of 2013, few have paid due attention to mid-level Russian defense and security experts who have warned of impending information confrontation with the West. Even the General Staff's normally diplomatic cyber-sages adopted a peace-through-the-knife approach, expressed in a journal article published as Wikileaks released a trove of DNC data in 2016:

> … the United States can enter into agreements with its geopolitical rivals only if they understand that they are opposed by an information potential as powerful as theirs. Therefore, the dialectic of interconnection and interdependence of political and military measures to counter the outbreak of war dictates the need to create a national information potential sufficient to deter possible aggression (Dylevskiy et al. 2016, 3–11).

That same year, a former deputy chief of the GRU discussed the "crisis" in relations between the West and Moscow against the mounting importance of information warfare, which, on a progressively greater scale, incorporated "cybernetic" operations that could achieve technical and psychological effects (Kondrashov 2016). Comprehending the specifics that guide Russian actors responsible for cyber and information operations can better prepare Western interlocutors and policymakers for managing a threat that will almost certainly exist throughout the near-term future.

4. CONCLUSION

Throughout the past few years, Russia's conceptualization of warfare has shifted to incorporate non-military means alongside armed violence. This transformation is exemplified by the increased relevance of information warfare in Russian doctrine.

According to this doctrine, information warfare consists of cyber and information operations and is an integral element of modern conflict. When discussing information warfare, official doctrine depicts Russia as a state nobly adhering to a defensive posture in an environment characterized by aggressive adversaries. The writings of Russian military scientists, however, illustrate an evolving interest in developing cyber weapons due to their effectiveness, appropriateness within the framework on contemporary conflict, and affordability. These analyses of offensive cyber tools seem more accurately aligned with the actual Russian practice of cyber and information operations that developed in parallel to Russia's thinking of contemporary conflict.

The actors and agencies involved in Russia's cyber operations evolved alongside Russia's perception of modern warfare and the threats posed by Western use of information technologies to further its military and foreign policy goals. In the first decades of the post-Soviet period, the FSB had a primary role in conducting cyber operations alongside the support of independent Russian hackers. Around the same time, a consensus formed among Russia's elite that warfare includes military and non-military measures during peace and wartime, and Russia's Defense Ministry increased its efforts to establish an organized and centrally controlled cyber force. These changes, coupled with the operational opportunities presented by Russia's intervention in Ukraine, enabled the GRU to adopt a leading position in offensive cyber operations, bringing a historical penchant for risk-taking and aggression to its operations. Additionally, the GRU's traditional command of information operations provided a natural place for cyber alongside information operations – the two core components of information warfare. These realities further enabled the transformation of Russia's strategic cyber operations from seemingly ad-hoc activities to more organized and centrally controlled campaigns that complement Russia's view of modern warfare.

Russia's conceptualization of information warfare and the units executing these operations are likely to drive future Russian cyber policy and strategy. The notion, for instance, that Russia faces aggressors who are utilizing evolving information communications technology to undermine Russia's military potential and society will almost certainly endure through the immediate future. At the same time, the idea that Russia's enemies are just as vulnerable to information means that Russia will probably safeguard the role of cyber and information operations within Russian doctrine and within the security and military organizations responsible for executing them for years to come. Although Russia's military inarguably will continue to value conventional assets and invest in modern warfighting technology, the growing prominence of unconventional means, particularly digital ones, in its ongoing competition with the West suggests that these capabilities will garner further attention in military doctrine, the writings of Russian military scientists, and state policy. It is possible that Russia's

leadership may choose to formally include research, development, and use of cyber weapons as an official line in its information warfare doctrine. However, this scenario seems unlikely considering that the current defensive nature of Russia's information warfare doctrine may enhance Russian claims of plausible deniability when being accused of conducting offensive cyber operations.

REFERENCES

Academy of Sciences. 1962. *Listovki otechestvennoy voyny* [Leaflets from the Patriotic War]. Moscow: U.S.S.R. Academy of Sciences.

Adamsky, Dmitry. 2015. "Cross-Domain Coercion: The Current Russian Art of Strategy." *Proliferation Papers* 54 (November):1–43.

Andrew, Christopher and Vasili Mitrokhin. 1999. *The Sword and The Shield: The Mitrokhin Archive and The Secret History of The KGB*. New York: Basic Books.

Antonov, S. I. et al. 2018. "*Modely ugroz sovmestnykh informatsionno-tekhnicheskikh i informatsionno-psikhologicheskikh vozdeystviy v gibridnykh voynakh* [Threat Models of Contemporary Information-Technical and Information-Psychological Impacts in Hybrid Wars]." *Informatsionnye Voyny* 2, no. 46:2–5.

Antonovich, P. I. 2011. "*O sushchnosti i soderzhanii kibervoyny* [On the Essence and Content of Cyber War]." *Voennaya Mysl*', no. 7:39–46.

Argumenty Vremeni. 2018. "*Osobyy front* [Special front]." October 1, 2018. https://svgbdvr.ru/voina/osobyi-front.

Balforth, Tom. 2018. "Putin Praises Skills of GRU Spy Agency Accused of UK Poison Attack." *Reuters*, November 2, 2018. https://www.reuters.com/article/us-britain-russia-putin/putin-praises-skills-of-gru-spy-agency-accused-of-uk-poison-attack-idUSKCN1N71YV.

Bazylev, S. I. et al. 2012. "*Deyateln'nost' Vooruzhennykh Sil Rossiyskoy Federatsii v informatsionnom prostranstve: printsipy, pravila, mery doveriya* [Activities of the Armed Forces of the Russian Federation in the Information Space: Principles, Rules, Confidence Building Measures]." *Voennaya Mysl*', no. 6:24–28.

Blank, Stephen. 2017. "Cyber War and Information War a la Russe." Carnegie Endowment for International Peace. October 16. https://carnegieendowment.org/2017/10/16/cyber-war-and-information-war-la-russe-pub-73399.

Burenok, V. M., ed. 2018. *Kontseptsii perspektivnogo oblika silovykh kompontentov voennoy organizatsii Rossiyskoi Federatsii* [Concepts of the Perspective Appearance of the Power Components of the Military Organization of the Russian Federation]. Moscow: Russian Academy of Missile and Artillery Sciences (RARAN).

Burtsev, M. I. 1981. *Perelom. Prozreniye* [Fracture. Insight]. Moscow: Voyennoye Izdatel'stvo.

Chekinov, Sergey and Sergey Bogdanov. 2013. "*O haraktere i soderhanii voiny novogo pokoleniia* [On the Character and Contents of the New Generation War]." *Voennaya Mysl*', no. 10:13–24.

Chekinov, Sergey and Sergey Bogdanov. 2015a. "*Voennoe iskusstvo na nachal'nom etape XXI stoletiya: problemy i suzhdeniia* [Military art in the initial stage of the XXI century: problems and judgments]." *Voennaya Mysl*', no. 1 (January):34–45.

Chekinov, Sergey and Sergey Bogdanov. 2015b. "*Prognozirovanie kharaktera i soderzhaniia voin budushchego: problemy i suzhdeniia* [Predicting the character and content of a future warrior: challenges and judgments]." *Voennaya Mysl*', no. 10 (October):41–9.

Cheshuin, S. A. 2009. "*Osobennosti sovremennogo informatsionnogo protivoborstva i ikh uchod pri podgotovke spetsialistov zarubezhnoy voyennoy informatsii v voyennom universitete* [The Features of Modern Information Confrontation During the Training of Specialists of Foreign Military Information at the Military University]." http://www.milpol.ru/sgs/sgs.html.

Chivvis, Christopher. 2017. "Hybrid War: Russian Contemporary Political Warfare." *Bulletin of the Atomic Scientists*, (August):316–21. http://www.tandfonline.com/doi/abs/10.1080/00963402.2017.1362903?journalCode=rbul20;

Connell, Michael and Sarah Vogler. 2017. *Russia's Approach to Cyber Warfare*. Arlington: Center for Naval Analysis.

Council on Foreign Relations. 2008. Connect the Dots on State-Sponsored Cyber Incidents-Agent.btz. November. https://www.cfr.org/interactive/cyber-operations/agentbtz.

Digital Forensics Research Lab. 2018. "#TrollTracker: Russia's Other Troll Team." *Medium*, August 2, 2018. https://medium.com/dfrlab/trolltracker-russias-other-troll-team-4efd2f73f9b5.

Diresta, Renee and Shelby Grossman. 2019. *Potemkin Pages & Personas: Assessing GRU Online Operations, 2014-2019*. Stanford: Stanford Internet Observatory.

Doffman, Zak. 2019. "Russian Secret Weapon Against U.S. 2020 Election Revealed in New Cyberwarfare Report." *Forbes*, September 24, 2019. https://www.forbes.com/sites/zakdoffman/2019/09/24/new-cyberwarfare-report-unveils-russias-secret-weapon-against-us-2020-election/#68503ec468f5.

Dylevskiy, I. N., et al. 2016. "*O dialektike sderzhivaniya i predotvrashcheniya voyennykh konfliktov eru* [On the Dialectic of Deterrence and Prevention of Military Conflicts of the Era]." *Voyennaya Mysl*', no. 7 (July):5–13.

Faizova, Liana, et al. 2018. "*12 Khakerov GRU: v chem SShA obvinili ofitserov Rossiyskoy Voennoy Razvedki* [12 GRU Hackers: Of What the United States Accused the Russian Military Intelligence Officers]." *The Bell*, July 13, 2018. https://thebell.io/ssha-obvinili-12-ofitserov-gru-vo-vzlome-pochty-demokratov-v-2016-godu/.

Galeotti, Mark. 2018. "I'm Sorry for Creating the 'Gerasimov Doctrine'." *Foreign Policy*, March 5, 2018. https://foreignpolicy.com/2018/03/05/im-sorry-for-creating-the-gerasimov-doctrine/.

Gavrilov, Yuriy. 2013. "*Ryadovoy Khaker* [Private hacker]." *Rossiyskaya Gazeta*, July 11, 2013. https://rg.ru/2013/07/10/roty-site.html.

Gazeta.ru. 2006. "*Kak Rossiya borolas's «Kavkaz-tsentrom*» [How Russia fought the Kavkaz Center]." March 9, 2006. https://www.gazeta.ru/2006/03/09/oa_191473.shtml.

Gerasimov, Valery. 2013. "*Tsennost' Nauki v Predvidenii* [The Value of Science is in Foresight]." *Voyenno Promyshlennyy Kuryer*. February 26, 2013. http://vpk-news.ru/articles/14632.

Gerasimov, Valery. 2014. "On the Role of Military Force in Contemporary Conflicts." In *Conference Proceedings*, III Moscow Conference on International Security. Moscow: Ministry of Defense of the Russian Federation. https://eng.mil.ru/files/MCIS_report_catalogue_final_ENG_21_10_preview.pdf.

Gerasimov, Valery. 2019. "*Vektory razvitiya voennoy strategii* [Vectors for the Development of Military Strategy]." Red Star, March 4, 2019. http://redstar.ru/vektory-razvitiya-voennoj-strategii/.

Giles, Keir. 2011. "Information Troops – A Russian Cyber Command?" In *3rd International Conference on Cyber Conflict*, edited by C. Czosseck, E. Tyugu, and T. Wingfield, 45–60. Tallinn: CCD COE.

Giles, Keir. 2016. *Handbook of Russian Warfare*. Research Division, NATO Defense College. November.

Greenberg, Andy. 2017. "Petya Ransomware Epidemic May Be Spillover from Cyberwar." *Wired*, June 28, 2017. https://www.wired.com/story/petya-ransomware-ukraine/.

Greenberg, Andy. 2019. *Sandworm: A New Era of Cyberwar and the Hunt for the Kremlin's Most Dangerous Hackers*. New York: Doubleday.

Greenberg, Andy. 2020. "The US Blames Russia's GRU for Sweeping Cyberattacks in Georgia." *Wired*, February 20, 2020. https://www.wired.com/story/us-blames-russia-gru-sweeping-cyberattacks-georgia/

Habr.com. 2013. "*O nauchnykh rotakh programmistov* [On the Science Units of Programmers]." *Khabr*, August 16, 2013. https://habr.com/ru/post/285590/.

Harris, S. and B. Devlin. 2019. "U.S. Investigating Sci-Hub Founder." *Washington Post*, December 20, 2019.

Haslam, Jonathan. 2015. Near and Distant Neighbors: *A New History of Soviet Intelligence*. New York: Farrar, Strauss, and Giroux.

Iasiello, Emilio J. 2017. "Russia's Improved Information Operations: From Georgia to Crimea." *Parameters* 47:2.

Ignatius, David. 2017. "Russia's Radical New Strategy for Information Warfare." *Washington Post*, January 18, 2017. https://www.washingtonpost.com/blogs/post-partisan/wp/2017/01/18/russias-radical-new-strategy-for-information-warfare/.

Izvestiya. 2016. "*Informatsionnoye protivoborstvo otrabotali na «Kavkaze-2016»*. Video [Information Confrontation Worked out in Caucasus-2016. Video]." September 14, 2016. https://iz.ru/news/632393.

Jensen, Benjamin, Brandon Valeriano, and Ryan Maness. 2019. "Fancy Bears and Digital Trolls: Cyber Strategy with a Russian Twist." *Journal of Strategic Studies* 42, no. 2:212–34.

Jonsson, Oscar. 2019. *The Russian Understanding of War*. Washington, DC: Georgetown University Press.

Kahn, David. 1996. *The Code-Breakers: The Comprehensive History of Secret Communications from Ancient Times to the Internet*. New York: Scribner.

Kari, Martti J. 2019. *Russian Strategic Culture in Cyberspace: Theory of Strategic Culture – A Tool to Explain Russia's Cyber Threat Perception and Response to Cyber Threats*. JYU Dissertations 122. University of Jyvaskyla. Faculty of Information Technology. October.

Kari, Martti J. and Katri Pynnöniemi. 2019. "Theory of Strategic Culture: An Analytical Framework for Russian Cyber Threat Perception." *Journal of Strategic Studies*, no. 11 (September):1–29.

Kavanaugh, Camino. 2015. "The UN GGE on Cybersecurity: The Important Drudgery of Capacity Building." *Council on Foreign Relations Blog*, April 2015. https://www.cfr.org/blog/un-gge-cybersecurity-important-drudgery-capacity-building

Kiselev, V. and A. Kostenko. 2015. "*Kibervoyna kak osnova gibridnoy operatsii* [Cyberwar as the Basis of Hybrid Operations]," *Armeiskii sbornik* 257, no. 11 (November):3–6.

Kofman, Michael. 2016. "Russian Hybrid Warfare and Other Dark Arts." *War on the Rocks*, March 11, 2016. https://warontherocks.com/2016/03/russian-hybrid-warfare-and-other-dark-arts/.

Kompromat.ru. 2002. "'*Utro' i Tsentr Zarubezhnoy Voyennoy Informatsii i Kommunikatsii Ministerstva Obrony Rossiyskoy Federatsii Predstavlyayut 'Chechnya: dokumental'noye kino*' ['Morning' and the Center for Foreign Military Information and Communications of the Ministry of Defense of the Russian Federation Present 'Chechnya: documentary']." November 6, 2002. http://www.compromat.ru/page_12433.htm.

Kondrashov, Vyacheslav Viktorovich. 2016. "*Informatsionnoe protivoborstvo v kiberneticheskom prostranstve* [Information Confrontation in the Cybernetic Space]." Scientific-Research Center of National Security Problems. August. https://nic-pnb.ru/analytics/informatsionnoe-protivoborstvo-v-kiberneticheskom-prostranstve/

Kozlov, Sergey. 2010. *Spetsnaz GRU: bezvremen'ye 1989-1999* [GRU Special Forces: Timelessness 1989–1999]. Moscow: Russkaya Panorama.

Kozlov, Sergey. 2013. "'*Zima-Leto 1942 Goda.' Spetsnaz GRU: istoricheskiye predposylki sozdaniya Spetsnaza 1941-1945* ['Winter-Summer of 1942.' GRU Special Forces: Historical Background of the Creation of the Special Forces 1941–1945]", vol. 2. Moscow: Russkaya Panorama.

Krikunov, A. 2011. "*Kiberprostranstvo vedushchikh gosudarstv v kontekste sovremennykh vyzovov i ugroz* [Cyberspace of Leading States in the Context of Contemporary Challenges and Threats]" *Morskoy Sbornik* 11 (November): 32-37.

Kritukov, Evgeniy. 2018. "*Kak SShA nashli 'sotrudnikov GRU', 'vmeshavshikhsya v vybory*' [How the United States found 'GRU officers', 'interfering in the elections']." *Vzglyad*, July 16, 2018. https://vz.ru/politics/2018/7/16/932761.html.

Kuznetsov, Sergey, Vasily Anisimov, Sergey Teslya, Igor Morozov. 2018. "*Kiberoperatsiya kak vid boyevykh deystviy* [Cyber Operations as a Kind of Mililtary Action]". *Zashchita i bezopasnost'* 1, no. 84:5.

Larin, D. A. 2017. *Kriptograficheskaya sluzhba Rossii: ocherki istorii* [Cryptographic Service of Russia: Essays on History]. Helios ARV.

Livejournal.com. 2015. "*Den' innovatsiy Ministerstva Oborony Rossii* [Innovation day of the Ministry of Defense of Russia]." October 6, 2015. https://bmpd.livejournal.com/1505576.html.

Medvedev, Sergei. 2015. "Offense-Defense Theory Analysis of Russian Cyber Capability." California: Naval Postgraduate School. https://pdfs.semanticscholar.org/19e3/ca12d73661182bd2a9e34dc2df81634deacf.pdf.

Mikryunov, V. Yu. 2015. "*Kak protivostoyat' agressii SSHA* [How to Resist U.S. Aggression]." *Vestnik Akademii voennykh nauk* 51, no. 2: 116–23.

Ministry of Defense of the Russian Federation. 2011. "*Kontseptual'nye vzglyady na deyatel'nost' Vooruzhennykh Sil Rossiyskoy Federatsii v informatsionnom prostranstve* [Conceptual Views on the Activities of the Armed Forces of the Russian Federation in the Information Space]." http://ens.mil.ru/science/publications/more.htm?id=10845074@cmsArticle.

Ministry of Defense of the Russian Federation. 2018. "*Prikazanie OPr/ 156* [Order OPr/156]." March 1. https://www.omgtu.ru/general_information/faculties/faculty_of_transport_oil_and_gas/deanary/news/2018/%D0%AD%D0%A0%D0%90%2009032018.pdf

Ministry of Foreign Affairs of the Russian Federation. 2000. "National Security Concept of the Russian Federation." January 10. https://www.mid.ru/en/foreign_policy/official_documents/-/asset_publisher/CptICkB6BZ29/content/id/589768.

Ministry of Foreign Affairs of the Russian Federation. 2011. "*Konventsiya ob obespechenii mezhdunarodnoy informatsionnoy bezopasnosti (kontseptsiya)* [Convention on International Information Security (Concept)]." September 22. https://www.mid.ru/foreign_policy/official_documents/-/asset_publisher/CptICkB6BZ29/content/id/191666.

Ministry of Foreign Affairs of the Russian Federation. 2013. "*Kontseptsiya vneshnoi politiki Rossiyskoi Federatsii* [Foreign Policy Concept of the Russian Federation]." February 12. http://www.mid.ru/foreign_policy/official_documents/-/asset_publisher/CptICkB6BZ29/content/id/122186.

Morris, Lyle J. et al. 2019. *Gaining Competitive Advantage in the Gray Zone: Response Options for Coercive Aggression Below the Threshold of Major War*. Santa Monica: RAND Corporation. https://www.rand.org/pubs/research_reports/RR2942.html.

Moscow-Russia.ru. n.d. "*Voyennyy Universitet Ministerstva Oborony Rossiyskoy Federatsii* [Military University of the Ministry of Defense of the Russian Federation]." http://moscow-russia.ru/voennyy-universitet-ministerstva-oborony/.

Moscow State Budgetary General Education Institute, School 1517. 2017. "*Soglashenie o sotrudnichestve v oblasti obrazovaniya* [Agreement on cooperation in the area of education]." May 19, 2017. https://1517.mskobr.ru/files/soglashenie_o_sotrudnichestve_v_oblasti_obrazovaniya_1517_akademiya.pdf

Nakashima, Ellen. 2017. "Inside a Russian Disinformation Campaign in Ukraine in 2014." *Washington Post*, December 25, 2017. https://www.washingtonpost.com/world/national-security/inside-a-russian-disinformation-campaign-in-ukraine-in-2014/2017/12/25/f55b0408-e71d-11e7-ab50-621fe0588340_story.html.

Nakashima, Ellen. 2018. "Russian Military Was behind 'NotPetya' Cyberattack in Ukraine, CIA Concludes." *Washington Post*, January 13, 2018. https://www.washingtonpost.com/world/national-security/russian-military-was-behind-notpetya-cyberattack-in-ukraine-cia-concludes/2018/01/12/048d8506-f7ca-11e7-b34a-b85626af34ef_story.html.

Nauchnaya Rota REB. 2015. "This Is a Recruiting Video for the Military Science Unit Dedicated to Electronic Warfare." *YouTube*. July 8, 2015. https://www.youtube.com/watch?time_continue=2&v=XoaR_0iANVA&feature=emb_logo.

Nezavisimaya Gazeta. 2000. "*Doktrina informatsionnoy bezopasnosti Rossiyskoy Federatsii* [Doctrine of Information Security of the Russian Federation]." September 15, 2000. http://www.ng.ru/politics/2000-09-15/0_infodoctrine.html.

Newsru.com. 2002. "*Tomskie khakery 3 goda vedut informatsionnuyu voynu protiv chechenskikh ekstremistov* [Hackers from Tomsk conducted an information war against Chechen extremists for three years]." January 30, 2002. https://www.newsru.com/russia/30Jan2002/hakery.html.

Parshin, S. and N. Bashkirov. 2019. "*Kiberugrozy i mezhdunarodnaya stabilnost*' [Cyberthreats and International Stability]." *Zarubezhnoe voennoe obozrenie*, no. 11 (November):3–10.

President of Russia. 2010. "*Voyennaya doktrina Rossiyskoy Federatsii* [Military Doctrine of the Russian Federation]." February 5. http://kremlin.ru/supplement/461.

President of Russia. 2016. "*Ob utverzhdenii doktriny informatsionnoy bezopasnosti Rossiyskoy Federatsii* [On Approving the Doctrine of Information Security of the Russian Federation]." May 12. http://kremlin.ru/acts/bank/41460/page/1.

Putin, Vladimir. 2012. "Vladimir Putin: *Byt' sil'nymi: garantii national'noy bezopasnosti dlya Rossii* [Be Strong: National Security Guarantees for Russia]." Rossiyskaya Gazeta no. 35 (February): 5708. https://rg.ru/2012/02/20/putin-armiya.html

Repko, S.I. 1999. *Voyna i propaganda, XV-XX vv*. [War and Propaganda, 15th–20th Centuries] Moscow: Novosti.

Ren.tv. 2019. "*Oruzhie budushchevo: Putin osmotrel sekretnye obraztsy vooruzheniya na 'Armiya-2019*' [Weapons of the Future: Putin Inspected Secret Weapons at 'Army 2019']." June 27, 2019. https://www.youtube.com/watch?v=Dviw_oSN4Yg.

Romashkina, N. P. And A. B. Kildobskiy. 2015. "*Novye metody protivoborstva XXI veka* [New XXI Century Methods of Confrontation]." *Vestnik Akademii voennykh nauk*, no. 1:134–39.

Rossiyskaya Gazeta. 2009. "*Postanovlenie pravitel'stva Rossiyskoy Federatsii ot 10 marta 2009 g. N 221 Moskva 'O prisuzhdenii premiy pravitel'stva Rossiyskoy Federatsii 2008 goda v oblasti nauki i tekhniki*' [Resolution of the Government of the Russian Federation from March 10, 2009 No. 221, Moscow 'On Awarding Prizes of the Government of the Russian Federation in 2008 in the Field of Science and Technology']." No. 0 (4872). March 20, 2014. https://rg.ru/2009/03/20/premii-nauk-tech-dok.html.

Rossiyskaya Gazeta. 2014. "*Voennaya doktrina Rossiyskoy Federatsii* [Military doctrine of the Russian Federation]." no. 298 (6570). December 30, 2014. https://rg.ru/2014/12/30/doktrina-dok.html.

Saltykov,Yevgeniy. 2014. "*V Rossii sozdany kibervoyska* [Cyber Forces Created in Russia]." *Vesti.ru*, May 12, 2014. https://www.vesti.ru/doc.html?id=1573024.

Security Council of the Russian Federation. 2013. "*Osnovy gosudarstvennoy politiki Rossiyskoy Federatsii v oblasti mezhdunarodnoy informatsionnoy bezopasnosti na period do 2020 goda* [Fundamentals of State Policy of the Russian Federation in the Field of International Information Security for the Period until 2020]." http://www.scrf.gov.ru/security/information/document114/.

Selivanov, V. V. 2020. "*O kompleksirovanii sredstv i sposobov podgotovki asimmetrichnykh otvetov pri obespechenii voyennoy bezopasnosti* [On Integrating Means and Methods for Preparing Asymmetric Responses in Ensuring Military Security]," *Voennaia mysl*', no.1 (January):48–60.

Sengupta, Kim. 2018. "Russian Spy Agency GRU Responsible for International Cyberwar UK Government Says." *Independent*, October 4, 2018. https://www.independent.co.uk/news/world/europe/russia-gru-sergei-skripal-hacking-cyber-war-donald-trump-elections-a8567356.html.

Shevyakin, Aleksandr. 2014. *KGB: sistema bezopasnosti SSSR [KGB: USSR Security System]*. Moscow: Algoritm.

Shil'bakh, K, and V. Sventsitskiy. 1927. *Voennye Razvedki* [Military Intelligence]. Moscow: Military Typography Directorate.

Slatter, Raphael. 2018. "Russian Hacking Europe Russia U.S. News Russian Hackers Posed as IS to Threaten Military Wives." *AP News*, May 8, 2018. https://apnews.com/4d174e45ef5843a0ba82e804f080988f/Russian-hackers-posed-as-IS-to-threaten-military-wives.

Starodubtsev, Y. I., V. V. Bukharin, and S. S. Semyonov. 2012. "*Tekhnosfernaya voyna* [Technosphere war]." *Voyennaya Mysl*' 7: 22–31.

TASS. 2014. "*Istochnik v Minoborony: v Vooruzhennykh Silakh RF sozdany voyska informatsionnykh operatsiy* [Source in the Ministry of Defense: Information Operations Troops Created in the Armed Forces of the Russian Federation]." May 12, 2014. https://tass.ru/politika/1179830.

Thomas, Timothy. 2019. *Russian Military Thought: Concepts and Elements*. The MITRE Corporation. Arlington, VA. August.

Thomas, Timothy. 2010. "Russian Information Warfare Theory: The Consequences of August 2008." In *The Russian Military Today and Tomorrow: Essays in Memory of Mary Fitzgerald*, edited by Stephen Blank and Richard Weitz. Carlisle: U.S. Army War College: 265-99.

Tikk, Eneken and Kerttunen, Mika. 2018. *Parabasis Cyber-diplomacy in Stalemate*. Norwegian Institute of International Affairs.

Tribun. 2018. "*Stali izvestny dannye o voyskakh «psikhov» Rossii* [Data on the Psycho Troops of Russia Became Known]." February 6, 2018. https://tribun.com.ua/47273.

Troianovski, Anton and Ellen Nakashima. 2018. "How Russia's Military Intelligence Agency Became the Covert Muscle in Putin's Duels with the West." *Washington Post*, December 28, 2018. https://www.washingtonpost.com/world/europe/how-russias-military-intelligence-agency-became-the-covert-muscle-in-putins-duels-with-the-west/2018/12/27/2736bbe2-fb2d-11e8-8c9a-860ce2a8148f_story.html.

Turovsky, Daniil. 2016."*Rossiyskiye vooruzhennyye kibersily. Kak gosudarstvo sozdayet voyennyye otryady khakerov* [Russian Armed Cyber Forces like a State Create Military Hacker Units]." *Meduza*, November 7, 2016. https://meduza.io/feature/2016/11/07/rossiyskie-vooruzhennye-kibersily.

Turovsky, Daniil. 2018. *Vtorzheniye: kratkaya istoriya russkikh khakerov [Invasion: A Brief History of Russian Hackers]*. Moscow: Inviduum.

Turovsky, Daniil and Rothrock, Kevin. 2018. "'It's Our Time to Serve the Motherland' How Russia's War in Georgia Sparked Moscow's Modern-Day Recruitment of Criminal Hackers." *Meduza*, August 7, 2018. https://meduza.io/en/feature/2018/08/07/it-s-our-time-to-serve-the-motherland.

U.S. Cyber Command. 2018. "Achieve and Maintain Cyberspace Superiority. Command Vision for US Cyber Command. " June 14. https://www.cybercom.mil/Portals/56/Documents/USCYBERCOM%20Vision%20April%202018.pdf?ver=2018-06-14-152556-010.

U.S. Department of The Treasury. 2018. "Treasury Sanctions Russian Federal Security Service Enablers." June 11 https://home.treasury.gov/news/press-releases/sm0410.

Vorob'ev, I. and V. Kiselev. 2013. "*Kibervoyna* [Cyber War]," Armeiskii sbornik, no. 8 (August):33–4.

Znamensk Garrison Military Court. 2011. "*Obzor sudebnoy praktiki rassmotreniya voennymi sudami grazhdanskikh del v 2010 godu* [Review of Judicial Practice of the Military Courts Review of Civil Cases in 2010]." February 28. http://znamenskygvs.ast.sudrf.ru/modules.php?name=docum_sud&id=336.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Measuring the Fragmentation of the Internet: The Case of the Border Gateway Protocol (BGP) During the Ukrainian Crisis

Frédérick Douzet
Professor
GEODE
University Paris 8
Saint-Denis, France
douzet@univ-paris8.fr

Louis Pétiniaud
PhD Candidate
GEODE
University Paris 8
Saint-Denis, France
l.petiniaud@gmail.com

Loqman Salamatian
GEODE
University Paris 8
Saint-Denis, France
salamatianloqman@gmail.com

Kevin Limonier
Associate Professor
GEODE
University Paris 8
Saint-Denis, France
Klimonier02@univ-paris8.fr

Kavé Salamatian
Professor
GEODE
University of Savoy
Annecy, France
kave.salamatian@gmail.com

Thibaut Alchus
GEODE
University Paris 8
Saint-Denis, France
thibaut.alchus@gmail.com

Abstract: This paper presents the results of a year-long research project conducted by GEODE (geode.science), a multidisciplinary team made up of geographers, computer scientists and area specialists.

We developed a new methodology for mapping cyberspace in its lower layers (infrastructures and routing protocols) in order to measure and represent the level of fragmentation of the Internet in areas of geopolitical tensions using the Border Gateway Protocol (BGP). Our hypothesis was that BGP could be used for geopolitical reasons in the context of a large-scale crisis, leading to a further fragmentation of the Internet. We focused on the Ukrainian crisis.

BGP is a core protocol of cyberspace that connects the tens of thousands of autonomous systems (ASes) that compose the Internet. Based on a 35-year-old technology, this protocol is easy to manipulate to re-route Internet traffic or even to cut off entire regions (BGP hijacks). Our results show actions on BGP implemented right after the 2014 Maidan Revolution, when Russian forces took control of the Crimean Peninsula and started to back separatist forces in Eastern Ukraine. In both cases, Russian authorities and separatist forces modified BGP routes in order to divert the local Internet traffic from continental Ukraine – drawing a kind of "digital frontline" consistent with the military one. The study of Donbass and of the Crimean Peninsula leads to important methodological findings to (1) define and map digital borders at the routing level; (2) analyze the strategies of actors conducting actions via BGP; (3) categorize these strategies, from traffic re-routing to cutting-off entire regions for intelligence or military purposes; and (4) anticipate future uses for BGP manipulations by identifying strategic bottlenecks within the network.

Keywords: *cyberspace, Ukraine, BGP, Russia, Crimea, Donbass, autonomous systems*

1. INTRODUCTION

On December 23, 2019, Russia claimed to have successfully tested disconnecting its network from the global Internet in an attempt to run a domestic alternative. Months earlier, the country had announced that it considered briefly unplugging itself from the Internet to test its cyber defense. This took place as the law n°608767-7 on the creation of a "sovereign Internet" came into force in November,[1] requiring technical alterations to provide Russia with the ability to control the Internet access points at its borders and to continue operating its domestic network in the event that it was disconnected from the global Internet.

These initiatives demonstrate the depth of Russia's strategic reflection on the structure of its connectivity and on the geopolitical importance of data routing. They are part

1 Federal Law n°608767-7 "On information, information technologies and information defense," https://sozd.duma.gov.ru/bill/608767-7.

of a larger strategy developed by Russia to secure sovereign control over what the authorities perceive as their national network, a geopolitical representation best captured by the term RuNet, widely adopted in Russia, and embodied by national platforms like Yandex or Vkontakte, to designate the post-Soviet linguistic, ethnic and cultural subspace of the web. The RuNet has since been used by Russian authorities to promote the representation of a sovereign cyberspace (Limonier 2018).

This strategy is not unprecedented. In November 2019, Iran accomplished just that when it cut off most traffic from the global Internet while operating its domestic network fully. The architecture of connectivity had been purposely redesigned to allow selective censorship of international traffic by connecting Iran's network to the outside with only three operators controlled by the government, thus creating a huge domestic intranet (Salamatian et al. 2019).

These initiatives have triggered concerns inside the Internet governance community about the increasing fragmentation of cyberspace and the risks it poses for its security and stability, not to mention online freedom and human rights. The question we ask in this paper is: "How can we measure and represent the fragmentation of cyberspace?" This paper presents the results of a year-long research project conducted by GEODE (geode.science), a multidisciplinary team composed of geographers, computer scientists and area specialists. We have developed a new methodology to map cyberspace in its lower layers (infrastructures and routing protocols) in order to measure and represent the fragmentation of the Internet in areas of geopolitical tensions using the Border Gateway Protocol (BGP).

Efforts to map cyberspace have focused on the physical infrastructure of the Internet, which is composed of cables, servers and other physical equipment that are grounded in physical territory and which can easily be mapped with the traditional tools of political and physical geography (Dodge and Kitchin 2001; Musiani et al. 2016). Other efforts have also attempted to capture the overall data traffic (Faravelon, Frénot, and Grumbach 2016). In the 2010s, much attention was given to the informational layer of cyberspace in the wake of jihadist propaganda and manipulations of information during democratic elections, leading to innovative cartographies of social networks and of the modes of content propagation (Howard et al. 2018; Limonier 2017). The strategic dimension of the BGP architecture and data routing, however, has been given much less attention in the scientific literature.

Jesse Sowell illustrates the importance of the lack of a top-down central governance model and the emergence of bottom-up governance models for groups of network operators – Internet exchange (IXP) groups – and other actors (Sowell 2012). This is made possible through the use of the BGP by autonomous systems (ASes) to establish

connections and exchange information between each other. BGP determines the routes data take and has been leveraged in the past by stakeholders to route traffic through specific paths and control the flow of information (Feamster and Ramachandra 2006). A relative flattening of the Internet structure has been observed, resulting from the emergence of major content providers like Netflix, Google, Amazon, Akamai, etc., along with major cable providers such as Angola Cables,[2] Me-We-Se,[3] and even Google, which owns 8.5% of Submarine Cables Worldwide (Zimmer 2018), that maintain a large part of the Internet traffic inside their networks (Wong 2016); however, BGP still has a primary role especially at the international level. It has also been manipulated by countries in order to block access to some content, to exclude some users from the Internet, to hijack traffic from other countries, or attack other countries' infrastructures. Many studies have focused on the inherent fragilities of a routing system designed in 1989 (Vervier, Thonnard, and Dacier 2015; Butler et al. 2010). Additionally, several articles have explored the BGP strategies of several nation-states (Edmundson et al. 2018; Wählslich et al. 2012).

Our hypothesis was that BGP could be manipulated for geopolitical reasons in the context of a large-scale crisis, leading to a further fragmentation of the Internet. We decided to focus on the Ukrainian crisis for several reasons.

First, the Ukrainian crisis presents a unique example of recent and direct military, economic, identity and diplomatic confrontation with Russia in the context of a major territorial conflict in Europe. Crimea and the two self-proclaimed republics of Donetsk and Luhansk in East Ukraine are spatial entities with disputed sovereignty sitting at the intersection of territorial and digital rivalries of power. In that sense, Crimea in particular can be perceived as a laboratory for Russia's strategies of appropriation.

Second, the anarchic development of the Internet in Russia and Ukraine has led to an abundance of ASes in both states, which provides larger sets of data with a greater level of precision. Finally, Russia has recently been testing methods to develop sovereign control of its network – particularly its physical infrastructure – through the re-nationalization of data networks, such as the obligation made in 2015 to maintain the data of Russian citizens in the country (Limonier 2018). But at the same time, Russia enjoys a very rich network with multiple external connections and its actors have been nurtured in the libertarian culture of Internet pioneers (Ermoshina and Musiani 2017). More recently, Russian authorities have asserted a need to organize the RuNet single-handedly.[4]

2 Angola Cables have emerged as an important actor in maritime Internet cables by providing direct links from Africa to South America.

3 Major maritime cables between Europe, the Middle-East and Asia.

4 "Совбез России поручил создать «независимый интернет» для стран БРИКС RBC," November 28, 2017, accessed March 9, 2020, https://www.rbc.ru/technology_and_media/28/11/2017/5a1c1db99a794783ba546aca.

This paper offers an overview of the topology of the Ukrainian network and its level of complexity in 2019. Then, through a longitudinal analysis of BGP data since 2013, it demonstrates the marginalization of Donbass and the appropriation of Crimea in cyberspace and raises the question of strategies of control these disputed territories have been subjected to, thus revealing the success and limits of Russia's venture for sovereign control in cyberspace.

2. METHODOLOGY

A. What is an Autonomous System?

The Internet is a network of networks characterized by its lack of centrality. It results from the interconnection of approximately 92,000 nodes (as of August 2019) called autonomous systems. An autonomous system (AS) is itself a network that manages its internal routing, distributes IP addresses to its customers and defines its access policies. Data transiting through the Internet from one point of the world to another usually crosses several independent ASes (6 on average) (Leguay et al. 2005).

Autonomous systems vary greatly in size and importance. A basic taxonomy divides them into three categories – Tier 1, 2 and 3 – which form an arborescent and partly hierarchical network structure. The most common types of Tier 1 ASes are intercontinental backbone carriers – such as Level 3 or Telia – or large national Internet Service Providers (ISP) – such as AT&T (United States), Orange (France), Rostelecom (Russia). Tier 2 ASes are generally medium-sized providers operating on regional or local scales. Tier 3 ASes (or "stub domains") are smaller networks run by a single company or university.

AS numbers, along with the blocks of IP addresses (or "prefixes"[5]) they manage, are allocated by the five Regional Internet Registries (RIR), themselves answering to the Internet Corporation for Assigned Names and Numbers (ICANN), one of the most important regulatory bodies of the Internet today. The administrator – either private or public – of each autonomous system determines a routing policy for its AS, which involves deciding which ASes to establish connections with and the behavior of its external routers when receiving data to be forwarded.

B. Why is BGP Political?

Notably, the security aspect of BGP routing in cases of traffic hijack, i.e., a redirecting of the traffic through malicious network nodes, has already stirred awareness of the political dimensions of routing.

However, BGP is political in ways that have not been investigated as much.

5 Set of several contiguous IP addresses that an AS can then assign to its users or customers.

First, an AS administrator wishing to connect to the global Internet has to establish relationships with other autonomous systems already connected to the network. The relationship can be of two types: a customer-to-provider (i.e., commercial) relationship, with an Internet Service Provider for instance; or a "peering," where two ASes estimate that they share approximately the same amount of traffic and set up a non-monetary relationship that allows their customers to exchange traffic. Moriano et al. analyzed the economic dimension of routing decisions (Moriano, Achar, and Camp 2016). Despite these choices being economic in nature, they also bear a political dimension.

Second, AS administrators implement routing algorithms that decide which path the packets of data will take to reach a destination, depending on commercial or security criteria, as well as on geopolitical considerations. When an AS receives information about a new possible path to reach a specific IP address, it chooses whether to change the path according to its preferences or keep the existing one. These routing policies integrate the basic rules of BGP along with the preferences set by AS administrators to create complex algorithms (Van Beijnam 2002).

Third, ASes contribute to the production of territories (Painter 2010). Through their routing policies, they define the paths and therefore the shapes of cyberspace. They are also implemented on physical territories and play a crucial role in providing places and people with Internet access and services, contributing to the development of territories. This is particularly true of remote places that rely on a limited number of ASes in order to access the global Internet, thus creating a digital territory defined by the topology of a network dependent on a few specific ASes. At the local level, the structure of ASes is critical to the resilience of the network (Chiu et al. 2015) and can result from spatial power strategies of various actors, a form of topological power (Allen 2011). The interconnection between states' ASes helps us understand how some countries might exert influence on others through connectivity and what relationships of dependency may exist.

Finally, BGP has been conceived of without security in mind and is very easy to manipulate for malicious or strategic purposes, such as espionage, censorship, disconnection, traffic hijack or the obfuscation of cyber attacks (Butler et al. 2010). Bearing the risk of observed BGP hijacks that could have resulted in threats of large-scale data exfiltration, Benton and Camp (2016) have proposed using BGP filters to ensure that packets are not being routed through problematic jurisdictions.

The strategic dimension of BGP deserves empirical studies. But mapping BGP data is a tremendous challenge because of the highly dynamic nature of this system. Routers can fail or restart. External connections between autonomous systems change at a very

fast pace and are announced through constant updates. For instance, an AS managed by the Russian company Vimpelcom (AS 8402) was found to have generated over 95,000 updates in seven days, which is not an extreme number. In addition, an autonomous system can change its information any time: the AS number can be reallocated, or the administrator can change its physical address and relocate to a different country. Relationships between ASes and routing policies are, for the most part, confidential and one challenge in measuring the Internet is to develop inference techniques to guess the policies of network operators and their relationships.

Despite these caveats, we were able to collect and process data to infer and map the topology of the Ukrainian network and its evolutions. Our approach is fundamentally interdisciplinary and involves research and fieldwork by regional specialists in geopolitics combined with the methodologies of computer science and mathematics.

C. What Data Did We Collect and Use?

Not all peering and customer-to-provider relationships are announced publicly. Our cartography is therefore mostly based on inference data, as opposed to data collected directly from operators. The AS relation graphs we infer are known to be incomplete. In particular, BGP path filtering policies do not expose less-preferred paths that would be chosen if the preferred announced paths were not available (Gregori et al. 2012). For this reason, we need to cross and combine our data with other sources (such as active measurements and IXP membership datasets) in order to obtain a consistent view of the network that can be mapped. Yet even this limited and incomplete view of the full AS graph is enough to monitor major changes to the Internet structure in Ukraine.

We have developed a BGP observatory that generates, every minute, a snapshot of a real-time AS graph that contains approximately 89,000 nodes and 220,000 links obtained by processing up to 30 BGP flows – announcing possible paths through a series of ASes – coming from different routers across the network. We have used the largest source of publicly available BGP routing data in 2019, RouteViews,[6] and the RIPE Routing Information Service (RIS),[7] which aggregates BGP messages from BGP monitors at cooperating ASes. These snapshots allow for a continuous monitoring of the logical layer of cyberspace at the AS level. We have collected more than ten terabytes of snapshots of AS graphs for a period of over three years. The AS graphs are inferred using path updates advertised by the routers running BGP to update neighboring routing tables (Roughan et al. 2011; Salamatian, Kaafar, and Salamatian 2018).

6 "Routeviews", accessed March 9, 2020, http://www.routeviews.org/routeviews.

7 "Routing Information Service – RIS," RIPE, accessed March 9, 2020, https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris.

We used several datasets and statistical methods:

1. We used graphs from the BGP observatory to represent the connectivity between the individual network operators (AS). Using that observatory, we augmented the BGP announcements by adding relevant information like 1) the name associated with the AS, 2) the country where the AS was registered, 3) the number of IP address prefixes announced by the AS and 4) the number of times a connection has appeared on the routing table.

2. We used the Potaroo blog to get statistics about the number of prefixes and ASes associated with each country year after year.[8]

3. We gathered the *AS relationships* (Dimitropoulos et al. 2007) inferred by the Caida Research Center at University of California San Diego, which indicate the underlying economic forces that drive the evolution of the Internet topology and its hierarchy.

4. In addition, we collected latency data using the Atlas network provided by RIPE, which allows any Internet user to install a probe on their server that can then be used by any other user to launch precise measures of connectivity.

Based on this data, our ambition was to study the topology of the networks and its consistence with the evolution of the topography of the country in a context of large-scale geopolitical crisis. The topological approach is highly valuable for approaching the reticular space of non-contiguous, enclave or exclave territories and the strategies of actors to reach this territory across space (Painter 2010; Latour 1987, 2005). By focusing on the crucial aspects of connectivity, such as data transits and network properties, the topological approach helps mobilize relevant concepts, such as accessibility, inclusion, borders, disjunction, continuity, intersection, connection and nodality (degree to which a node is the point of convergence between different routes) (Severo and Venturini 2016). In a nutshell, the topological approach is "first and foremost a reduction of complexity in the name of representing more complexity" (Piper 2013).

D. Limitations of Our Methodology

The BGP view is well-known to be incomplete. In particular, peer-to-peer (p2p) links are known to be harder to observe than customer-to-providers (c2p) links (Gao 2001; Ager et al. 2012; Cohen and Raz 2006). A contribution of this paper is to show that even this incomplete view provides valuable geopolitical insights. Moreover, c2p links reflect real economic strains and are therefore better indicators of the power relationships that shape the topology of the network.

8 "BGP Routing Table Analysis Reports", Houston G. Blog, accessed March 9, 2020, https://bgp.potaroo.net/.

Another shortcoming of BGP analysis is usually caused by the incompleteness of available information on AS owners stored in the Whois registry, as well as the unreliability of IP geolocation databases at the regional and local levels (Poese et al. 2011). In this work, we compensate for these shortcomings through a qualitative analysis, based on OSINT (Open Source Intelligence). We use various sources of information to find geographical data on the most important actors of Ukraine's connectivity and their relationships to policymakers.

In addition, our analysis is based on the routes available for data traffic and not on the quantification of the actual volume of traffic that circulates through these routes, as we cannot access this level of granularity in BGP data. However, we are able to evaluate the importance of a link through the number of announced BGP paths, and the number of BGP prefixes that cross it. Although these values do not precisely give the amount of traffic, it allows us to understand how central a link is for the overall routing. Moreover, we consider all the ASes to be nodes, despite their diversity (governmental, private, universities, geographically bounded to cities, etc.).

Last but not least, we need to acknowledge BGP's intense fungibility and lack of fixed relationships: on average, more than 5,000 route changes happen every second in the whole Internet. Most of them result from operational constraints (like a router rebooting), but some of them are also caused by relationship changes between ASes. This is the reason why we track data overtime in order to be able to provide longitudinal studies and avoid over-interpreting isolated instances of routing changes. Nevertheless, BGP is a highly dynamic environment and no cartography could possibly pretend to be fully accurate and exhaustive.

3. STRUCTURE OF CONNECTIVITY IN UKRAINE

A. A Rich and Diverse Network

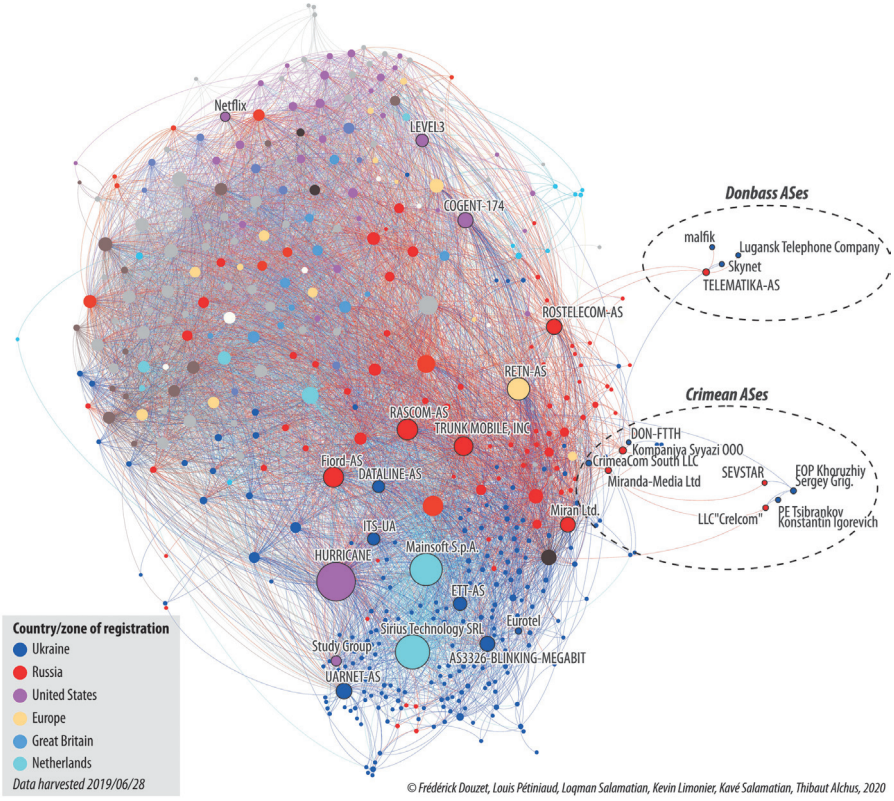
We first looked at the architecture of Ukraine's ASes and the way they are connected to the rest of the world. Our first graph (Figure 1) represents, as of June 26, 2019, all Ukrainian ASes and their immediate neighbors, meaning ASes that have a direct relationship with at least one Ukrainian AS. Each node represents an AS, and each link a relationship (commercial or peering). For clarity, we eliminated from the graph ASes with fewer than five neighbors and provided the name of significant ASes only.[9] The nodes are colored according to the country the ASes are registered in.[10] Although this information is often reliable, it can hide part of the reality. Large ASes that operate

9 The names of the ASes (including quotation marks, numbers and capital letters) are based on the RIPE database. They are the official names of the autonomous systems. As such, the names are based on the decision of the administration of each AS, and do not always match the name of their parent company.

10 "List of country codes and RIRs," RIPE, accessed March 9, 2020, https://www.ripe.net/participate/member-support/list-of-members/list-of-country-codes-and-rirs.

internationally are likely to change their country of registration for political reasons, as we will see below, hence the need for qualitative research for graph analysis.

FIGURE 1. REPRESENTATION OF UKRAINIAN AUTONOMOUS SYSTEMS AND THEIR DIRECT NEIGHBORS, JUNE 2019



The size of the nodes (ASes) in the graph depends on their betweenness (the state of being between) centrality, i.e., the proportion of the shortest paths between all nodes of the graph that go through this link. The betweenness centrality measures the impact of disconnecting a link for the global connectivity of the network (Ma et al. 2008) and points to the most important nodes in the routing architecture of a country.

Finally, we used Force Atlas 2, a visualization algorithm for a representation of our graph. This algorithm is based on a concept of repulsion – with nodes pushing each other away but links attracting nodes closer – simulates the dynamics of a physical system to spatialize the network. In other words, the closer the nodes, the more connections they share.

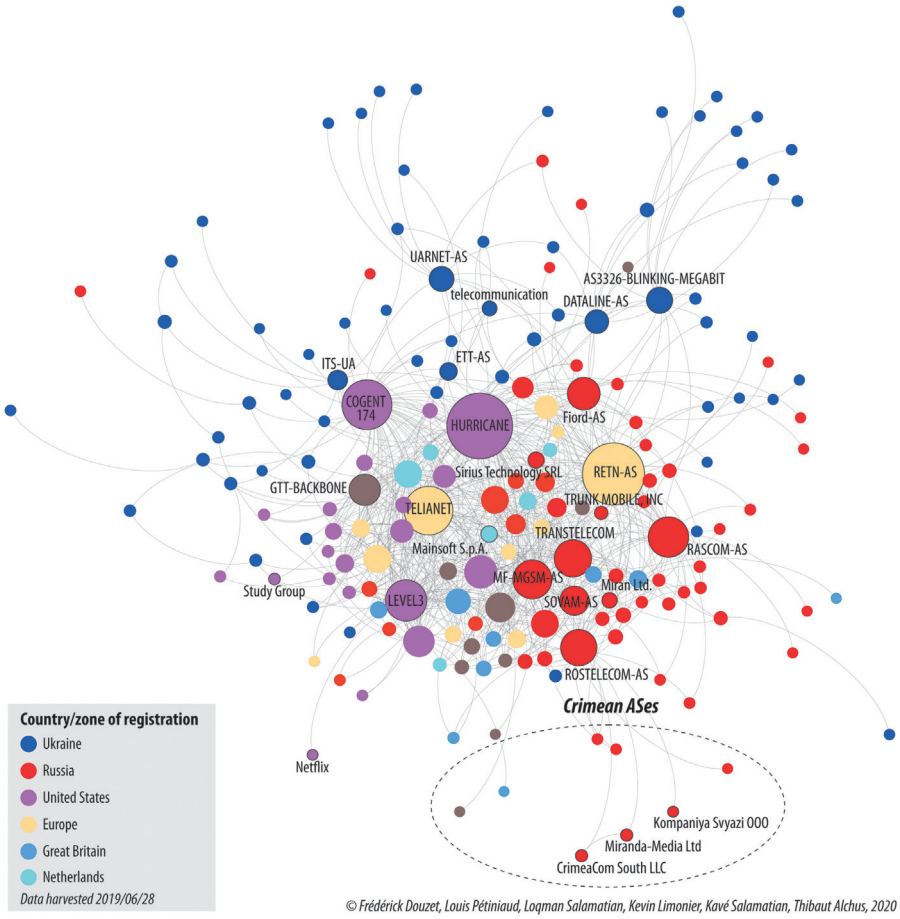
The graph shows that Ukraine possesses a very rich network, with nearly 2,200 allocated ASes, among which over 1,800 are active (i.e., announced in our BGP data). This profusion is characteristic of Ukraine and Russia, which counts 5,176 active ASes. Both countries aggregate a high proportion of ASes compared to their population: about 24,300 users per AS on average in Ukraine and 30,485 in Russia, compared to about 110,000 users per AS on average in other European countries. However, most of the Ukrainian ASes are of small size on the graph, which reflects their low centrality in the network, i.e., the fact that they do not have many neighbors and therefore do not attract much of the traffic. Most of them are stub ASes (Tier 3) and serve a limited, sometimes very small, area.

This disproportion has historical roots and can be explained by the relative anarchy in which the Internet was developed in the post-Soviet republics during the 1990s and 2000s while European countries were structuring their network around major historical telecom operators, such as France Telecom in France. This profusion is reinforced by the competition between multiple economic actors with diverging interests in a rather opaque system controlled by oligarchs (Limonier 2018). It makes the network particularly resilient, but also complex and difficult to control, as we will see below.

B. The Polarization of Ukraine's Cyberspace Between Russian and Western Routes

The Ukrainian network is clearly structured around two poles: Russia on the one hand, the United States on the other hand, along with a myriad of other (mostly European) countries. Two Italian ASes (Sirius and Mainsoft) are highly visible due to their aggressive peering policy, but are less relevant when looking closely at the results. The structure of the network offers a great diversity of paths to the global Internet, but they are under the control of either Russian ASes (Rostelecom, Rascom-AS, Transtelecom) or major American or European ASes (GTT, Level 3, Cogent, Hurricane). Therefore, the architecture of the network reflects the geopolitical situation of Ukraine: split between major powers.

FIGURE 2. SIMPLIFIED REPRESENTATION OF UKRAINIAN ASES AND THEIR NEIGHBORS, JUNE 2019



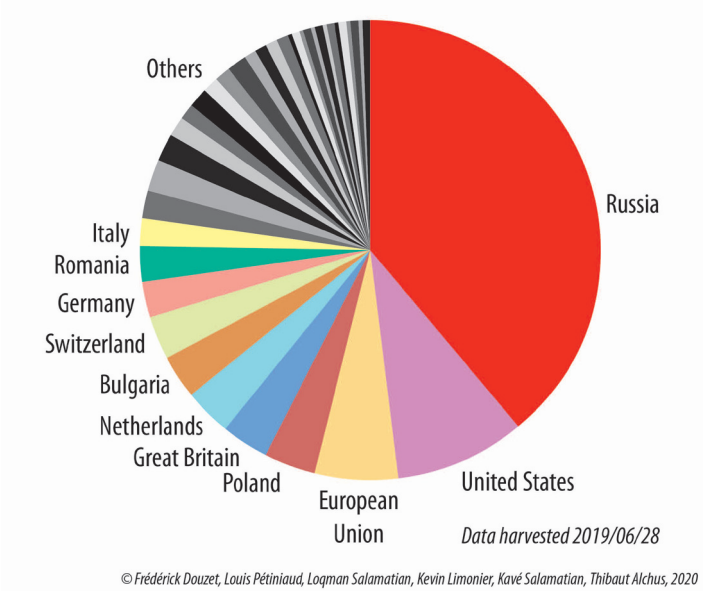
A simplified view of the network (Figure 2) gives a clearer understanding of who the major players are. In Figure 2, we only kept the Ukrainian ASes, their six most important neighbors and the links that appeared the most often in our routing table, thus eliminating 94% of the links, along with the less central ASes.[11] We can see that most Ukrainian ASes have disappeared due to their small size. Only the most important ASes remain in our graph, which are mainly foreign ones. The divide between the two poles is even more distinct.

The two Italian ASes are less central, which means that despite their many connections, they do not capture most of the traffic. They are fully integrated into the galaxy

11 We selected the top 6% of the links that appeared the most often in our routing table (i.e., more than 484 times).

of American and European ASes that connect the main Ukrainian ASes to smaller Ukrainian ASes and to foreign ASes of medium centrality. The UK, Germany and the Netherlands are important, yet usually not essential, points of transit. The place of RETN on the graph seems inconsistent, but is not surprising. Registered in Europe, RETN was once declared Ukrainian, but is currently administered by a major telecom company based in Saint-Petersburg; hence, the proximity to Russian ASes.[12]

FIGURE 3. DISTRIBUTION BY COUNTRY OF REGISTRATION OF UKRAINIAN ASES' NEIGHBORS, JUNE 2019



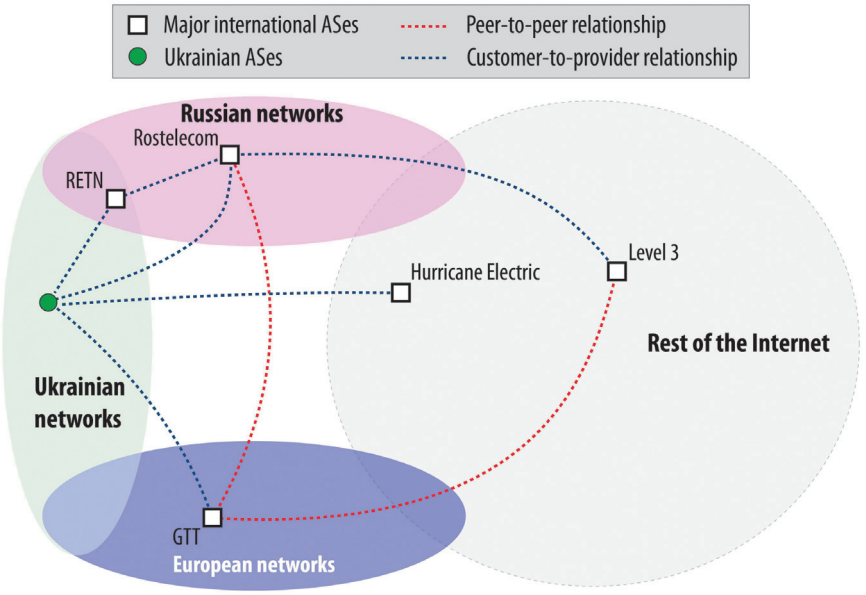
Overall, Russia occupies a major place in the graph, with 95 ASes connected to a Ukrainian AS, representing nearly 40% of all neighbor ASes (Figure 3). In comparison, the United States has only 22 ASes connected, but a number of them are major Tier 1 ASes. The American presence has strongly increased since June 2019 in our graph, with the direct connection of Hurricane Electric – a (near[13]) Tier 1 AS – to Ukrainian ASes. This observation might be explained by the strategic competition between the US and Russia over Ukraine, but could also be part of a wider phenomenon of centralization around major providers which are directly connected to smaller ASes without intermediaries. Some call this the "flattening of the Internet" (Böttger et al. 2019). It requires further investigation to confirm our hypotheses.

12 "RETN network map", RETN, accessed March 9, 2020, https://retn.net/networkmap/.

13 A Tier 1 network can reach every other network on the Internet solely via peering links. Hurricane Electric can reach "only" 85% of the Internet via peering links alone.

Last but not least, we notice on the margins of the graph a couple of clusters of ASes that represent the sub-spaces of Ukraine's cyberspace, namely Crimea and the Donbass regions, which are dealt with in the next section.

FIGURE 4. UKRAINE'S PATHS TO THE GLOBAL INTERNET, JUNE 2019



C. A Complex Network, Hard to Control

The Ukrainian network is therefore two-headed, with a few major ASes providing most paths toward the global Internet (Figure 4). It also appears to be rich and distributed from the heart of the country, with some peripheral ASes, on average two jumps away from a major Ukrainian AS. The disputed territories are exceptions, as seen below. Following the Berkman Center of Internet and Society, we measured the complexity score of the network (Roberts et al. 2011) to better understand its architecture. This metric captures the complexity of a network within a country by looking at the diversity in the announcements of IP addresses assigned to the country. A high complexity score means the possibility of a larger set of routing paths, through more providers, to connect ASes to each other or to the global Internet. A low complexity score (below 1) indicates with more certainty a network that is easy to control and to protect by periphery defense (like gatekeepers or firewalls). Also, high complexity means that it will be more difficult to introduce major changes, for example through a cyber-attack, into the structure of the country's network. In other words, changing

the structure of a complex network involves putting in a lot of effort to overcome the native resilience resulting from the complexity of the network.

The results of the complexity score and control value calculation (Table 1) show that both Russia (141) and Ukraine (79) have very high complexity scores compared to other countries of the region. This means that if important changes are observed in the connectivity structure of these two countries, this would likely result from a deliberate effort to implement such a transformation.

TABLE I: COMPLEXITY SCORE AND CONTROL VALUE IN THE BLACK SEA REGION, DECEMBER 2019

Countries	Number of ASes	Complexity Score	Control Value
Ukraine	1821	79.2	0.18
Russia	5049	141.3	0.10
Bulgaria	620	23.6	0.14
Turkey	448	2.7	0.06
Georgia	91	1.8	0.46
Romania	1040	39.8	0.41
Moldova	136	3.6	0.49

We calculated another metric proposed by the Berkman Center: the control value (Roberts et al. 2011). This metric leverages the notion of "points of control," defined as the minimal set of ASes needed to connect 90% of advertised IPs in the country to the external world. The lower the control value, the greater the centralization of the network (Salamatian et al. 2019).

Ukraine requires only 18% (about 328 ASes) of its total number of ASes to announce 90% of its allocated IP addresses. This means that controlling these 328 ASes could be enough to control almost all traffic, considering the small size of ASes. Although the control value is not very high, the profusion of ASes makes the network particularly complex and therefore difficult to control. Russia's control value is lower (10%), but the number of ASes and the complexity score are much higher.

Overall, Ukraine's network is diverse and very complex with a multiplicity of actors involved and a few powerful foreign neighbors who ensure most of the external paths. The strategies of territorial appropriation developed by Russia in Crimea and the development of geopolitical conflicts on the ground therefore constitute a major challenge in cyberspace.

4. THE FRAGMENTATION OF CYBERSPACE IN UKRAINE

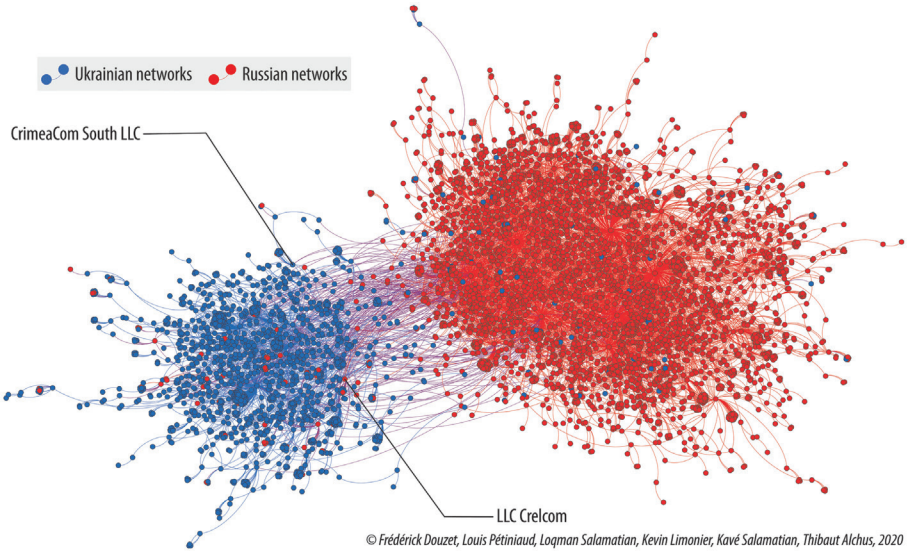
Crimea and Donbass have been forcefully fought over since 2014. Crimea was annexed in March 2014 and is now ruled by the Russian Federation, although Ukraine continues to claim sovereignty over the oblast. Russia controls the territory and its main infrastructures, including the Kertch bridge and supply channels for water, energy and Internet access. The two self-proclaimed republics of Donetsk and Luhansk are in a very different situation, since they pit separatists backed by unofficially involved Russian forces against Ukrainian military forces that have been joined by independent volunteers.

Although these conflicts are still active, the territorial limits have stabilized and the Ukrainian government has lost power in both territories; to Russia in Crimea and to independently elected bodies in Donbass. Network control is part of the territorial disputes that redefine power relationships. This process enhances the loss of sovereign control by Ukraine's government and reinforces the dependency of these territories on external actors.

A. The Emergence of Crimea and Donbass as Separate "Territories" in Cyberspace

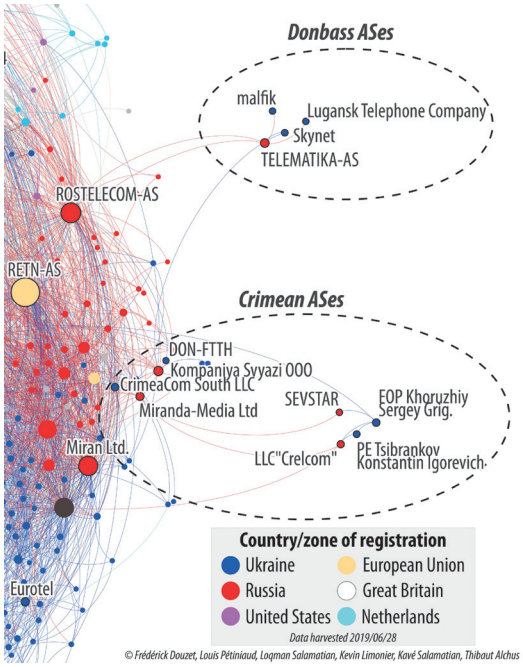
Our graphs reveal the fragmentation of Ukraine's cyberspace. In 2013, the ASes of Crimea and Donbass were fully integrated into the Ukrainian network (Figure 5), as illustrated by the central position of the Crimean ASes Crelecom and CrimeCom at the heart of Ukraine's network. There was no identifiable region in Ukraine's cyberspace. These ASes are indeed quite distant from each other on the graph, which means that they did not share the same connections. In 2019, we can see sub-regions clearly, characterized by clusters of ASes at the periphery of the networks, which reflect two different geopolitical situations: the annexation of Crimea by Russia, on the one hand, and the marginalization of Donbass in the wake of the separatist uprising, on the other hand.

FIGURE 5. REPRESENTATION OF UKRAINIAN AND RUSSIAN ASES, SEPTEMBER 2013



The close-up in our first graph (Figure 6) demonstrates the successful appropriation by Russia of Crimea's connectivity. Most Crimean ASes are now registered in Russia and are connected to Russian ASes. Russia managed to capture nearly all the traffic and there are almost no paths left to Ukraine's main ASes. Despite this amalgamation, Crimea remains at the periphery of the Russian network, as illustrated by the position of Crimean ASes on the graph. This spatial distance means that the number of connections between Crimean ASes and Russian ASes is limited. This observation could be explained by a deliberate strategy to isolate Crimea's network to better control it. This hypothesis requires further research.

FIGURE 6. DONBASS AND CRIMEA, "SCATTERED" TERRITORIES OF CYBERSPACE, JUNE 2019



Donbass sits in a different position, apparently at the interface of Crimea and Ukraine. Its ASes have clearly migrated toward Russia but still share many connections with Ukrainian ASes. Donbass has become marginalized in the Ukrainian network, but not fully integrated into the Russian network.

How did this happen? We chose to focus on the case of Crimea to uncover the steps that led to the split between Crimea and Ukraine's cyberspace.

B. Russia's Strategies of Territorial Appropriation of Crimea in Cyberspace

Russia demonstrated its will to control the network as early as February 28, 2014, when a Russian commando force seized the building and equipment of the Ukrainian company Ukrtelekom and cut its cables that linked Crimea to Ukraine, thus disconnecting the largest part of the peninsula from the Internet. But the complexities of Internet connectivity required a more sophisticated strategy to address the concerns of Russian officials, as expressed by the Prime Minister in a tweet on March 24, 2014: data transit between Crimea and Moscow could not be provided by foreign companies.

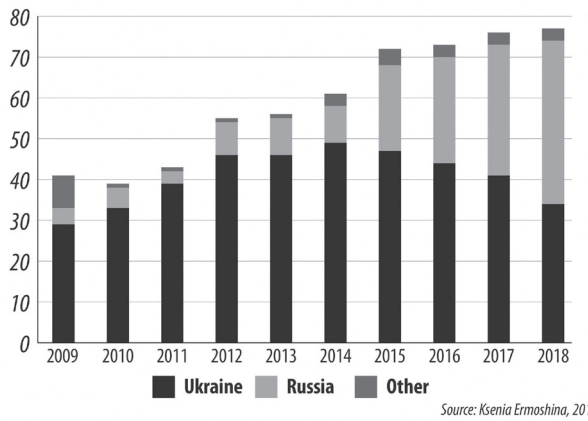
Russia restored access to the Ukrainian connectivity, but put in place a progressive strategy that led three years later to the digital annexation of Crimea and the

marginalization of Donbass. At that time, it was physically impossible for Russia to ensure that all the Crimean traffic could go directly to the Russian mainland through the Kerch stretch, which explains why access to Ukrainian connectivity had to be restored in 2014. For at least three years, Russia had to rely partially on two fiber-optic cables connecting Crimea to the rest of the world via the isthmus of Perekop, a wide strip of land connecting the peninsula with the Ukrainian mainland. To avoid this situation, Rostelecom, the Russian company in charge of its implementation, bought 1,700 km of cables from the two main ISPs, Datagroup and Atrakom, and unveiled 46 km of new cables through the Kertch Strait, the only option to avoid transit through Ukrainian hubs, on April 25, 2014.

Meanwhile, in mid-April 2014, Rostelecom invested 15 million rubles in one of its branches, Miranda Media, to run operations in Crimea[14] and the first cable was activated on July 17, to secure strategic military communications in priority. Miranda Media (AS201776) popped up in the routing tables. A second cable (905 km long) was deployed on May 15, 2017 to absorb the traffic of Internet users as Miranda Media became more central. In July meanwhile, Ukraine's government decided to stop providing Internet access to Crimea through the two optic cables that linked them together.[15]

In addition, Russian companies pursued an active strategy of buying local ISPs and convincing others to use their services. Many ISPs became Russian as a result of pressure, to avoid potential problems, or out of loyalty to the government (Ermoshina 2018). The following graph (Figure 7) shows the overtime evolution of registration of ASes in Crimea.

FIGURE 7. DISTRIBUTION OVER TIME OF CRIMEA'S ASES BY COUNTRY OF REGISTRATION



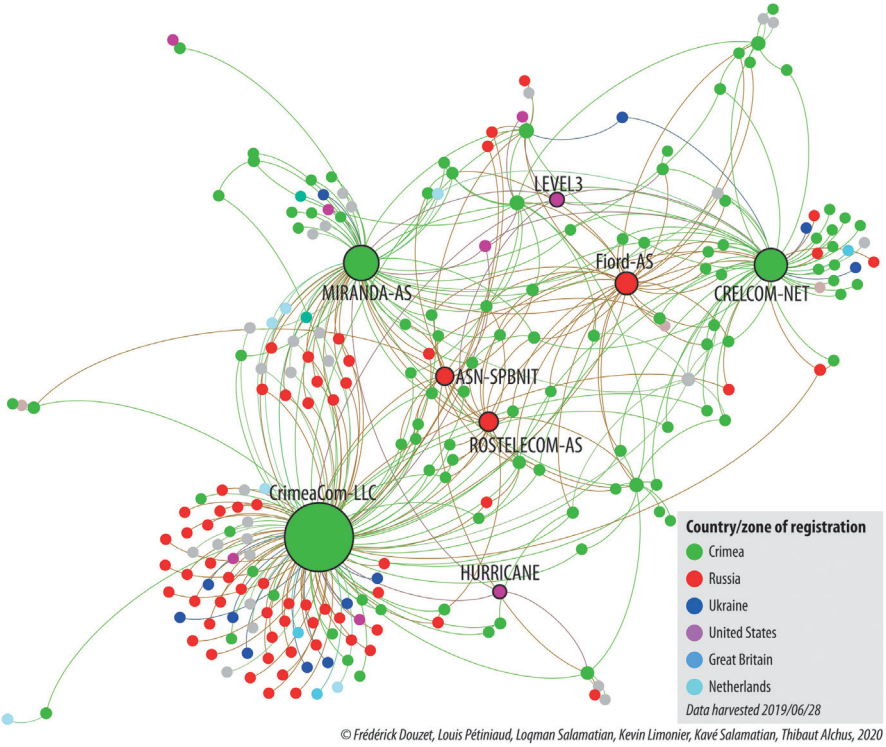
Source: Ksenia Ermoshina, 2018

14 ""Ростелеком" потратил $30 млн на покупку инфраструктуры в Крыму," *Comnews*, May 8, 2014, accessed June 6, 2019, https://bit.ly/2QOnpEw.

15 "Украина прервала связь с Крымом," *Comnews*, July 24, 2017, accessed June 6, 2019, http://www.comnews.ru/content/108850/2017-07-24/ukraina-prervala-svyaz-s-krymom.

This situation is also due to Ukrainian sanctions against companies that continued to provide Internet connectivity to Crimea after the annexation. Major ASes, Russian ones included, were forced to withdraw from Crimea to avoid jeopardizing their activities elsewhere. As a result, smaller Crimean ASes started growing bigger and more central in a network that became structured around three major ASes: Miranda Media (AS201776), Crelcom (AS6789) and CrimeaCom (AS28761), all registered in Russia. A graph of Crimean ASes and their direct neighbors (Figure 8) shows the centrality of these three providers in the network. At the heart of this graph are three major Russian ASes: Rostelecom, SPBNIT and Fiord. A few Tier 1 American ASes are present, but are not central in the graph (Hurricane, Level 3).

FIGURE 8. REPRESENTATION OF CRIMEAN ASES AND THEIR DIRECT NEIGHBORS, JUNE 2019



Crimea's network became increasingly centralized around three major actors close to the Russian power. Although we could not measure it, we can hypothesize that the level of complexity of Crimea has decreased as a result of these changes.

The amalgamation of Crimea with the Russian network is confirmed by a measure of latencies. We used the Atlas network to target two IP addresses, one in Simferopol (Crimea) and one in Nova Kakhovka (Kherson), and sent over 900 pings from Ukraine, Russia, Romania, Georgia, Moldova, Bulgaria and Belarus. The results presented in the two following maps (Figure 9 and 10) clearly show the difference of connectivity between these two points in the network: Crimea is in the privileged access zone of Moscow, no longer in Kiev's.

FIGURE 9. CRIMEA'S TOPOLOGICAL PROXIMITY TO MOSCOW, MEASURED BY LATENCIES, 2019

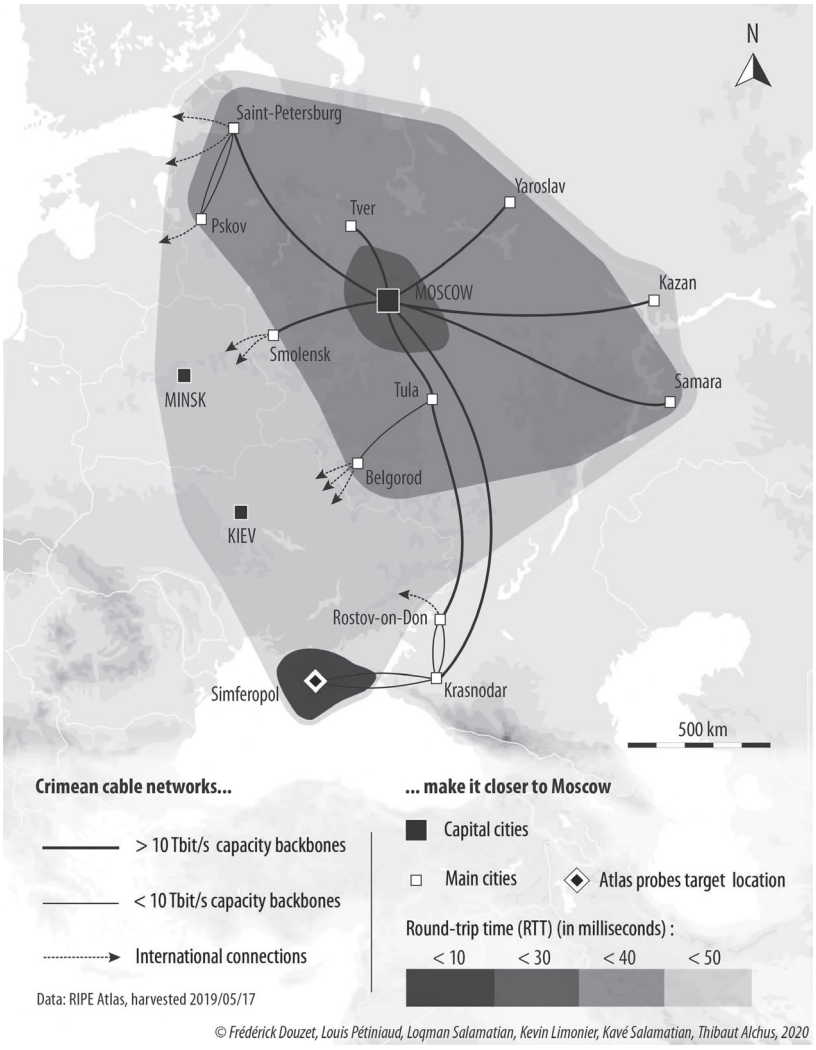
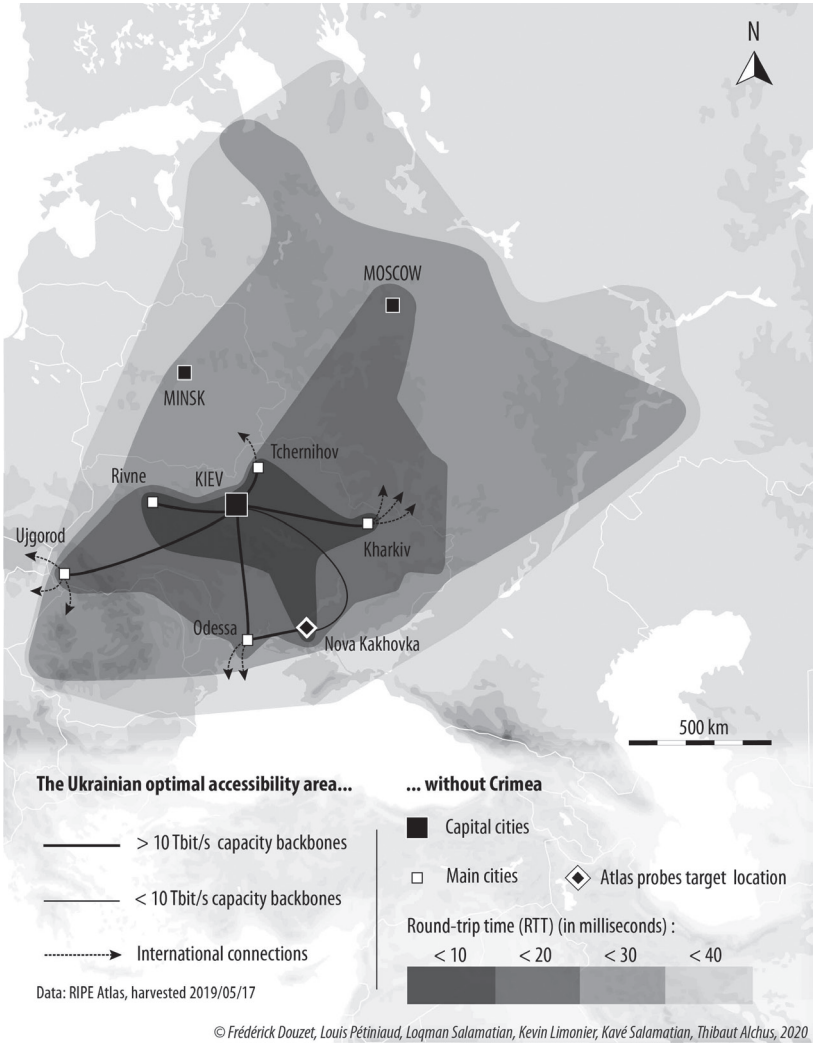


FIGURE 10. CRIMEA'S TOPOLOGICAL MARGINALIZATION FROM UKRAINE, MEASURED BY LATENCIES, 2019

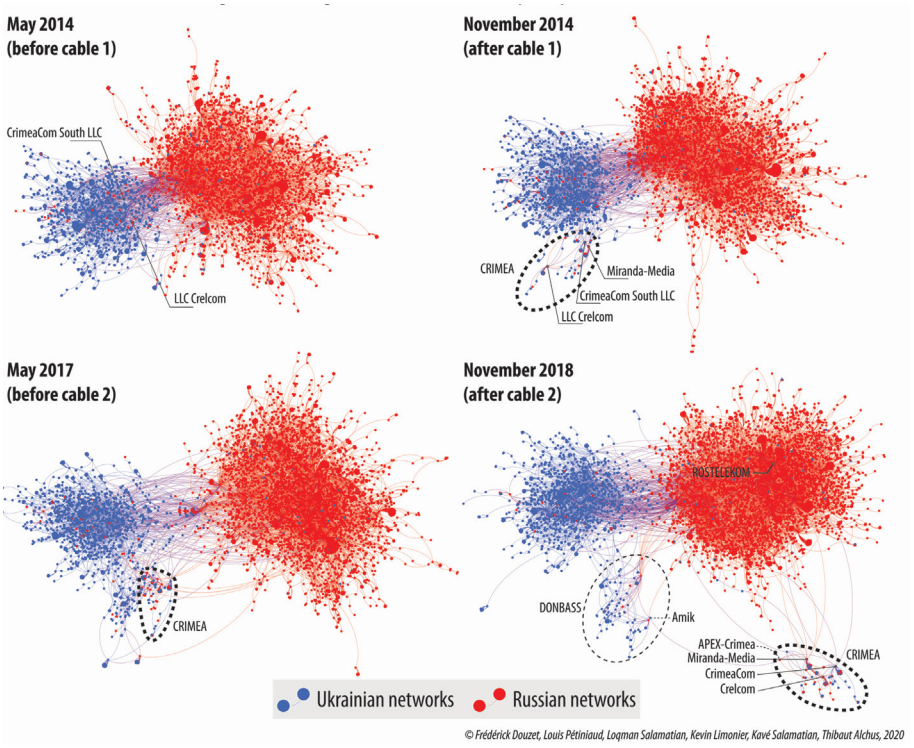


C. Longitudinal Study of Ukraine and Russia ASes

The fragmentation is also well reflected by the antagonism between Ukraine and Russia. A selection of graphs representing ASes of Russia and Ukraine at different times of the crisis show the clear relationship between the evolution of the topography and topology of Crimea and the unfolding geopolitical events (Figure 11). We observe three tendencies: 1. the break-up and progressive integration of Crimea into the

Russian network; 2. the marginalization of Donbass; 3. the gradual increase in the distance between the two countries.

FIGURE 11. THE FRAGMENTATION OF UKRAINE'S CYBERSPACE, 2014–2018



5. CONCLUSION

Our study shows that geopolitical conflicts over territories do have a clear impact on the shape of cyberspace, and that the same dynamics of annexation and fragmentation can be observed. In Crimea and Donbass, Russian authorities and separatist forces were able to attract digital traffic into their respective networks and modify BGP routes in order to divert the local Internet traffic from continental Ukraine, drawing a kind of "digital frontline" consistent with the military one. This resulted in the fragmentation of Ukraine's cyberspace, leading to the emergence of separate sub-spaces. The study of the Crimean Peninsula and of Donbass leads to important methodological findings that can allow us to: (1) define and map digital borders at the routing level; (2) analyze the strategies of actors conducting actions via BGP; (3) categorize these strategies, from

traffic re-routing to cutting off entire regions for intelligence or military purposes; and (4) anticipate future uses for BGP manipulations by identifying strategic bottlenecks within the network.

The ability to demonstrate a government's influence and deliberate strategies of territorial appropriation requires further work. Through the combination of BGP data and fieldwork-based research, we were able to demonstrate that the case of Crimea reveals a clear intent, on the part of Russia, to achieve a control of the connectivity in addition to the physical territory in the peninsula. This case study also reveals the role played by Ukraine in this dynamic of fragmentation through its decision to sanction companies providing connectivity to Crimea.

As a result, the cartography of routing paths should be seen as an additional tool to observe geopolitical conflicts, and their consequences on cyberspace, that should be used in combination with other methodologies to obtain a more complete picture.

ACKNOWLEDGMENTS

The authors thank Maxime Cherveaux for proof-reading this paper.

REFERENCES

Ager, Bernhard, Nikolaos Chatzis, Anja Feldmann, Nadi Sarrar, Steve Uhlig, and Walter Willinger. 2012."Anatomy of a large European IXP." In *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication, Helsinki, Finland, August 2012*, 163–174. New York, NY: Association for Computing Machinery. DOI: 10.1145/2342356.2342393.

Allen, John. 2011. "Topological Twists: Power's Shifting Geographies." *Dialogues in Human Geography* 1, no. 3: 283–298.

Benton, Kevin, and L. Jean Camp. 2016. "Firewalling Scenic Routes: Preventing Data Exfiltration via Political and Geographic Routing Policies." In *SafeConfig '16: Proceedings of the 2016 ACM Workshop on Automated Decision Making for Active Cyber Defense, 2016*, 31–36. DOI: 10.1145/2994475.2994477.

Böttger, Timm, Gianni Antichi, Eder L. Fernandes, Roberto di Lallo, Marc Bruyere, Steve Uhlig, Gareth Tyson, and Ignacio Castro. 2019. "Shaping the Internet: Ten Years of Internet Growth." arXiv:1810.10963v3.

Butler, Kevin, Toni R. Farley, Patrick McDaniel, and Jennifer Rexford. 2010. "A Survey of BGP Security Issues and Solutions." In *Proceedings of the IEEE, Chennai, India, January 2010* 98, no. 1, 100–122. https://ieeexplore.ieee.org/abstract/document/5357585/.

Chiu, Yi-Ching, Schlinker Brandon, Radhakrishnan Abhishek Balaji, Katz-Bassett Ethan, and Govindan Ramesh. 2010. "Are We One Hop Away from a Better Internet?" In *Proceedings of the 2015 Internet Measurement Conference, Tokyo, Japan, October 2015*. 523–529. DOI: 10.1145/2815675.2815719.

Cohen, Rami, and Danny Raz. 2006. “The Internet Dark Matter-on the Missing Links in the AS Connectivity Map.” In *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications, 2006*. DOI: 10.1109/INFOCOM.2006.234.

Dimitropoulos, Xenofontas, Dmitri Krioukov, Marina Fomenkov, Bradley Huffaker, Young Hyun, George Riley, and Kimberly C. Claffy. 2007. “AS Relationships: Inference and Validation.” *ACM SIGCOMM Computer Communication Review* 37, no. 1: 29–40. DOI: 10.1145/1198255.1198259.

Dodge, Martin, and Robert Kitchin. 2003. *Mapping Cyberspace*. London: Routledge.

Edmundson, Anne, Roya Ensafi, Nick Feamster, and Jennifer Rexford. 2018. “Nation-State Hegemony in Internet Routing.” In *COMPASS ‘18: Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, Menlo Park, United States, June 2018*. DOI: 10.1145/3209811.3211887.

Ermoshina, Ksenia. 2018. “A Routing Interregnum: Internet Infrastructure Transition in Crimea after Russian Annexation.” *35th Chaos Communication Congress* (35C3). DOI: 10.5446/39274.

Ermoshina, Ksenia, and Francesca Musiani. 2017 “Migrating Servers, Elusive Users: Reconfigurations of the Russian Internet in the Post-Snowden Era.” *Media and Communication* 5, no. 1. DOI: 10.17645/mac.v5i1.816.

Faravelon, Aurélien, Stéphane Frénot, and Stéphane Grumbach. 2016. “Chasing Data in the Intermediation Era: Economy and Security at stakes.” *IEEE Security and Privacy Magazine*, Part 2, 14, no. 3: 22–31. DOI: 10.1109/MSP.2016.50.

Feamster, Nick and Anirudh Ramachandra. 2006. “Understanding the Network-level Behavior of Spammers.” *ACM SIGCOMM Computer Communication Review* 36, vol. 4: 291. DOI: 10.1145/1151659.1159947.

Gao, Lixin. 2001. “On Inferring Autonomous System Relationships in the Internet.” *IEEE/ACM Transactions on Networking* 9, no. 6: 733–745. DOI: 10.1109/90.974527.

Gregori, Enrico, Alessandro Improta, Luciano Lenzini, Lorenzo Rossi, and Luca Sani. 2012. “On the Incompleteness of the AS-level Graph: A Novel Methodology for BGP Route Collector Placement.” In *IMC ‘12: Proceedings of the 2012 Internet Measurement Conference, Boston, MA, United States, November 2012*, 253–264. DOI: 10.1145/2398776.2398803.

Howard, Philip N., Bharath Ganesh, Dimitra Liotsiou, John Kelly, and Camille François. 2018. “The IRA, Social Media and Political Polarization in the United States, 2012-2018.” *Project on Computational Propaganda*, Oxford Internet Institute.

Latour, Bruno. 1987. *Science in Action*. Washington, DC:Howard University Press.

Latour, Bruno. 2005. *Reassembling the Social*. Oxford: Oxford University Press.

Leguay, Jeremie, Matthieu Latapy, Timur Friedman and Kave Salamatian. 2007. “Describing and Simulating Internet Routes”. *Computer Networks, Elsevier Science*, vol. 51, n°8: 2067-2085.

Limonier, Kevin. 2017. “Guerre hybride russe dans le cyberespace.” *Hérodote* 166–167, no. 3: 145–163. DOI: 10.3917/her.166.0145.

Limonier, Kevin. 2018. *Ru.Net: Géopolitique du cyberespace russophone*, Paris: L’Inventaire.

Ma, Nan, Jiancheng Guana, and Yi Zhao. 2008. “Bringing PageRank to the Citation Analysis.” *Information Processing & Management* 44, no. 2: 800–810. DOI: 10.1016/j.ipm.2007.06.006.

Moriano, Pablo, Soumya Achar, and L. Jean Camp. 2016. “Macroeconomic Analysis of Routing Anomalies.” In *TPRC 44: The 44th Research Conference on Communication, Information and Internet Policy, Arlington, VA, United States, September-October 2016*. https://ssrn.com/abstract=2755699.

Musiani, Francesca, Derrick L. Cogburn, Laura DeNardis, and Nanette S. Levinson. 2016. *The Turn to Infrastructure in Internet Governance*. Houndmills: Palgrave Macmillan.

Painter, Joe. 2010. "Rethinking Territory." *Antipode*, 42, no. 5: 1090–1118.

Piper, Andrew. 2013. "Reading's Refrain: From Bibliography to Topology." *ELH* 80, no. 2: 388.

Poese, Ingmar, Steve Uhlig, Mohamed Ali Kaafar, Benoit Donnet, and Bamba Gueye. 2011. "IP Geolocation Databases: Unreliable?" *ACM SIGCOMM Computer Communication Review* 41, no. 2: 53–56. DOI: 10.1145/1971162.1971171.

Roberts, Hal, David Larochelle, Rob Faris, and John Palfrey. 2011. *Mapping Local Internet Control*. Harvard: Berkman Klein Center, Harvard Center for Internet & Society. https://cyber.harvard.edu/netmaps/mlic.pdf.

Roughan, Matthew, Walter Willinger, Olaf Maennel, Debbie Perouli, and Randy Bush. 2011. "10 Lessons from 10 Years of Measuring and Modeling the Internet's Autonomous Systems." *IEEE Journal on Selected Areas in Communications* 29, no. 9. DOI: 10.1109/JSAC.2011.111006.

Salamatian, Loqman, Dali Kaafar, and Kavé Salamatian. 2018. "A geometric approach for Real-Time Monitoring of Dynamic Large Scale Graphs. AS-Level Graphs Illustrated." Paper given at *ACM SIGCOMM Internet Measurement Conference, Boston, MA, United States, October-November 2018*.

Salamatian, Loqman, Frédérick Douzet, Kavé Salamatian, and Kevin Limonier. 2019. "The Geopolitics Behind the Routes Data Travels: A Case Study of Iran." arXiv:1911.07723.

Severo, Marta, and Tommaso Venturini. 2016. "Enjeux topologiques et topographiques de la cartographie du Web." *Réseaux*, 195, La Découverte.

Sowell, Jesse H. 2012. "Empirical Studies of Bottom-Up Internet Governance." In *Proceedings TPRC, Arlington, VA, September 2012*.

Van Beijnum, Iljitsch. 2002. *BGP: Building Reliable Networks with the Border Gateway Protocol*. O'Reilly Media, Sebastopol, CA, United States.

Vervier, Pierre-Antoine, Olivier Thonnard, and Marc Dacier. 2015. "Mind Your Blocks: On the Stealthiness of Malicious BGP Hijacks." In *Proceedings 2015 Network and Distributed System Security Symposium, San Diego, CA, United States, February 2015*. DOI: 10.14722/ndss.2015.23035.

Wählisch, Matthias, Thomas C. Schmidt, Markus de Brün, and Thomas Häberlen. 2012. "Exposing a Nation-Centric View on the German Internet - A Change in Perspective on AS-Level." *In: Taft N., Ricciato F. (eds) Passive and Active Measurement. PAM 2012. Lecture Notes in Computer Science, vol 7192. Springer, Berlin, Heidelberg*.

Wong, Joon Ian. 2016. "How Streaming Video Changed the Shape of the Internet." *Quartz*, October 5, 2016. https://qz.com/742474/how-streaming-video-changed-the-shape-of-the-internet/.

Zimmer, Jameson. 2018. "Google Owns 63,605 Miles and 8.5% of Submarine Cables Worldwide." *BroadbandNow*, September 12, 2018. https://broadbandnow.com/report/google-content-providers-submarine-cable-ownership/.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Cyber in War: Assessing the Strategic, Tactical, and Operational Utility of Military Cyber Operations

Matthias Schulze
Associate
International Security Division
German Institute for International and Security Affairs (SWP)
Berlin, Germany

Abstract: The study analyzes the use of cyber capabilities in war and conflict situations. The research question is: What good is cyber in war? What is the utility of military cyber operations in conflict situations and what obstacles exist? The paper analyzes a small set of cases where cyber capabilities have been used for military purposes. Using the 'three levels of warfare' heuristic, the study outlines the potentials and operational restrictions of military cyber operations. The analysis proposes a set of variables and hypotheses, such as the timing of use of cyber capabilities and the operational complexity of a cyber operation, for further theory building.

Keywords: *cyber in war, military cyber operations, levels of war, strategic cyber attacks, tactical cyber, small-n case study*

1. INTRODUCTION

North Korea's leader, Kim Jong-un, allegedly heralded cyber capabilities as an "all-purpose sword" that guarantees "ruthless striking capability" (Young Kong, Gon Kim, and Lim 2019). Popular books, such as *The Perfect Weapon* by David Sanger, frame cyber capabilities as the Swiss Army knife of war, which can be used for all

kinds of purposes. Offensive cyber capabilities are often seen as "force multipliers" with high precision, global-reach, relatively low cost and potentially high impact (Smeets 2018b, 98). Strategic cyber warfare conducted by the military to shut down an adversary's critical infrastructure, a type of "cyber Pearl Harbor," has been hyped as the next revolution in military affairs, but has not materialized so far (Lawson 2013). Besides deterrence, norms, and taboos, one explanation for this lack of cyber warfare could be the severely limited strategic utility of cyber in war (Libicki 2009, 117). Beyond the strategic level, more and more studies highlight the limitations of cyber operations in conflict situations.

This paper aims to analyze the utility and potential unsuitability of military cyber operations in war or conflict contexts. For that purpose, the study analyzes a small set of cases where cyber operations have been used for military purposes. The paper uses the 'three levels of warfare' heuristic, which distinguishes between cyber operations on the strategic, operational, and tactical levels, to sketch out the utility of cyber technology on each of these. This approach is taken because prior research suggests that the strategic utility of cyber is limited (Valeriano, Jensen, and Maness 2018); the same, however, may not be true for the other levels of warfare. The research question thus is: What good is cyber technology in war? What is the utility of military cyber operations in conflict situations and what obstacles exist?

Cyber warfare often lacks the central components of war: large-scale physical destruction, massive violence and compelling an actor to the political will of another (Rid 2012). While the stand-alone use of cyber capabilities might not be regarded as war, the use of *cyber in war* is a feature of almost all modern armed conflicts, from Kosovo 1998 to Ukraine 2014. This study focuses on the use of cyber in war, generally understood as military cyber operations that are defined as a "sequence of coordinated actions with a defined military purpose in cyberspace; requiring cyber capabilities" (van Haaster 2019, 148). The term *operations* indicates a sequential or parallel use of offensive cyber attacks in a coordinated manner, in contrast to singular cyber attacks. The goal of the conduct of war, in general, is not just to destroy or disable physical infrastructures and forces, but to achieve psychological effects, such as compelling an enemy to do one's will (Clausewitz 1982).

2. TYPES OF CYBER OPERATIONS IN WAR

Cyber in war has the following characteristics. First, cyber attacks in war are often conducted by military organizations, such as cyber commands. Second, these are often, but not exclusively, targeted against opponents' military infrastructures such as headquarters, command and control, and weapon systems. Cyber attacks in war

are often counter-force attacks and stand in contrast to the use of strategic cyber attacks in peacetime to influence the decision calculus of adversaries (Smeets 2018b). Third, they serve a military rather than an intelligence purpose, such as supporting other forces in combat, and thus have a military intention. The distinction between military and non-military cyber operations is, however, not clear-cut. Ambiguities remain because of the attribution problem, as well as the functional overlap with cyber espionage that is conducted by intelligence agencies.

Military theory divides war into three levels: strategic, operational, and tactical. These levels are interrelated and what happens on one level influences the others. The strategic level deals with issues of "how to win a war" (Bateman 2015). The strategic level allocates national resources and instruments of power to achieve victory in war. Strategies ideally define how to use the various means of state power, including cyber capabilities, toward the end of achieving peace. As Clausewitz famously highlighted, the political level of war often cannot be clearly separated from the strategic (Clausewitz 1982). In most democracies, the political level – that is, elected politicians not generals – decide when to go to war.

Strategic attacks, whether kinetic or cyber, are those that try to achieve strategic objectives such as weakening the adversaries' ability or will to engage in conflict (US Air Force 2019). Strategic cyber attacks typically target sources of national power or society in general (Libicki 2009, 117). John Arquilla defines strategic cyber warfare as a "means of striking in very costly, disruptive ways at an adversary without a prior need to defeat opposing military forces in the field, at sea, or in the air" (Arquilla 2017). Strategic cyber attacks are often used as a stand-alone capability that can be executed without mobilizing other, more conventional forces. Cyber attacks in peacetime that target vital functions of a state, such as critical infrastructures, fall into the strategic category, but so do the defend forward or preparation of the battlefield strategies. Strategic cyber attacks are also used for tacit-bargaining, coercion or deterrence (Borghard and Lonergan 2017).

Below the strategic level is the operational level, which is often concerned with the conduct of campaigns and the question of how to employ forces in various theaters, such as a geographic region (Valeriano and Maness 2015, 243). The goal on the operational level is to obtain advantages over the enemy in a series of battles. Targets on the operational level tend to be military, such as enemy ships, tanks or troops, especially if battles take place far from civilian infrastructures. The Allied invasion of Normandy, Operation Overlord, was one operation among many in a specific theater of war to achieve the strategic objective of defeating Nazi Germany. Operational cyber attacks often serve as an "adjunct" function to traditional military forces (Gartzke 2013, 66), that is using the cyber domain together (jointly) with the other domains of

warfare (Sanger 2018, 41). For instance, cyber capabilities can be used as a distraction in the early phases of a war to sow confusion and panic on one front, while other forces move in from another direction unobstructed (Jun, LaFoy and Sohn 2015, 15).

Lastly, the tactical level is the realm of combat engagements between individual warfighters and units in a combat situation. Most traditional weapon systems operate at this level. The tactical level thus deals with the conduct and movement of troops in a given terrain. Not much has been written about *tactical cyber*. Cyber operations on the tactical level take place "in the context of a traditional kinetic battlefield, where authorization, deconfliction, and control for the specific operation is at battalion level or lower" (Metcalf and Barber 2014). Deconfliction means overcoming different areas of responsibility of military command levels or between agencies, for example between high-level intelligence agencies and battalion units. An example could be a combat mission, such as a hostage rescue in a "smart city", where video cameras are hacked to provide special forces with situational awareness (Crane and Peeke 2016). IT equipment, drones or GPS devices of combatants could be interfered with using cyber operations. Tactical cyber could take two forms. One is the integration of IT specialists into small units in the field. The second variant is that soldiers can rely on "remote cyber support" from a unit placed somewhere at a safe distance (Porche et al. 2017, 47).

3. CASE STUDIES

This paper uses an inductive or hypothesis-generating case study design that is not based on a full-fledged theory (Levy 2008, 5). Since there is no statistically meaningful number of cases of military cyber operations, the study examines variables in a small number of cases (small-n). The purpose is to develop theoretical propositions about the use of military cyber operations which then can be tested by future research. The aim is to deduce variables that explain the utility of cyber in war. The cases in question are known instances where cyber operations were used in a military context or were conducted by a military organization such as a cyber command. The study excludes the use of cyber capabilities for political or economic espionage, as well as instances of cybercrime, for methodological reasons. Focusing on military operations also excludes non-state actors, which makes the research more manageable.

A. Strategic Level

Many inter-state cyber operations happen at the strategic level. Most of them are intentionally designed to stay below the threshold of an armed attack to avoid escalation into conventional conflict (Valeriano and Maness 2015, 183). To this date, there is no case of a coordinated, strategic cyber war campaign against another state

that reached the level of an armed attack or could be easily classified as war. The closest to this is the planned operation Nitro Zeus, by US Cyber Command against Iran, that was uncovered in 2016. This was a contingency plan in case its predecessor, operation Olympic Games, better known as Stuxnet, and diplomatic efforts to limit Iran's nuclear program, failed. According to David Sanger, the plan included striking at Iran's air defense, transportation, and communications systems, as well as crucial parts of the power grid (Sanger and Mazzetti 2016). The pre-emptive attack would almost certainly have affected civil critical infrastructure in peacetime. Nitro Zeus was a large-scale effort involving thousands of intelligence personnel who placed backdoor implants in Iranian computer networks, preparing the battlefield. Insiders describe it as "a huge, expensive undertaking, beyond the reach of anyone but a few nation-states" (Sanger 2018, 45). Like Stuxnet, it probably required years of preparation, reconnaissance, simulation and malware testing. The plan was never executed, and one can only speculate as to why.

Fear of retaliation in the context of vulnerability of US critical infrastructures is certainly one explanation. Iran's cyber corps attacked US financial institutions after Stuxnet was uncovered (Sanger 2018, 46). Such a large-scale strategic attack would most likely be regarded as the use of force in international law and thus likely escalate into a conventional conflict in the region. Additionally, with complexity comes uncertainty about the reliability of implants that must remain undetected by adversaries for some time. Then there is the risk of collateral damage. In addition to these issues, Nitro Zeus clearly shows one benefit of strategic cyber warfare, and that is having another option on the table, in case negotiations break down (Smeets 2018b, 97). Press reports are unclear whether Nitro Zeus was conceived as a stand-alone, strategic operation that would shut down Iran's system "without firing a shot" and thus without risking the lives of US troops in a probably lengthy war (Sanger and Mazzetti 2016). It also could have been conceived as a pre-emptive first strike of a more conventional conflict. Both options are conceivable.

In academia, skeptics argue that even in war, the strategic utility of military cyber operations is limited. Martin Libicki maintains that strategic cyber attacks cannot be used effectively for two key elements of war, namely permanently disarming or degrading enemy conventional forces or occupying and holding a territory (Libicki 2009, 59). A central issue of cyber in war is that strategic cyber capabilities are target-dependent, in that they need to be tailored to specific target configurations. Because malware must be custom-built, it is more difficult to have stockpiles that are up-to-date once conflict occurs. Tomahawk missiles are built once and are ready to use during their expected shelf-life of 30 years (Defense Industry Daily 2020); but 0-day vulnerabilities have a shorter life cycle and shelf-life and they cannot be stored in the same way (Ablon and Bogart 2017). Therefore, 0-day malware must be

written beforehand to be operational once fighting breaks out. Therefore, wiping-out an entire country with strategic cyber attacks requires a concerted and simultaneous effort of different attack vectors that need to be prepared and maintained in advance. This requires a huge logistical effort of keeping track of the status of implants and especially how different attack vectors are intertwined or depend on each other. High-value targets, such as critical infrastructures and command and control systems, are often air-gapped and require specialized intelligence to gain access. In many instances, this requires time-consuming social engineering in advance to gain a foothold on a system. This implies high operational complexity for a vast-scale strategic attack that permanently disrupts another country over a sustained period. Since the damage of cyber attacks is often temporary and reversible, additional resources need to be continuously spent to shut down a nation permanently (Smeets 2018a). This reduces the strategic utility of cyber capabilities in war (Borghard and Lonergan 2017, 477) and suggests that strategic cyber attacks may be valuable only in the early stages of a conflict, for example, to generate surprise effects. Cyber attacks tend to be most effective when they are not expected (Kostyuk and Zhukov 2017). In the early stages of a conflict, malware arsenals are stacked up and 0-day vulnerability arsenals are not yet burned. The longer a conflict lasts, and the longer cyber barrages endure, the fewer available 0-day vulnerabilities should remain and the lower the expected utility of cyber operations.

Another argument against the utility of strategic cyber operations comes from research on strategic air-power. Proponents of strategic bombardments of cities in war argue that pain inflicted on the adversary's population will help to turn it against its government, thereby reducing the enemy's will to resist. Empirical studies find that strategic air-raids against civil infrastructures rarely produce this effect. In contrast, attacks against civil infrastructures are often perceived as illegitimate. Instead of reducing the enemies' will to resist, they inflict anger and create a rally-around-the-flag-effect, where the population moves to support the war efforts of its government (Pape 1996). Reasoning by analogy, the same might be true for a military cyber operation that shuts down an entire nation (Lawson 2013, 94–95).

One generally assumed advantage of strategic cyber operations is that they provide military planners with a flexible instrument that can be adjusted to the specific target. Max Smeets argues that, like a covert operation, they provide state leaders with an alternative option to act without necessarily risking escalation into a physical conflict (Smeets 2018b, 97). In times when there are only bad options available, cyber solutions might be the lesser evil, because, if used cautiously, they provide states with plausible deniability and an alternative to conventional strikes or the deployment of special forces. Strategic cyber attacks can be designed to create only temporary and reversible effects; they might provide a non-lethal option as well. Reversible damage

might be an option for more humane conduct of war, but risking enemy recovery might not be in the military's interest. In situations of doubt, shooting a missile and permanently destroying a military target seems to be preferable to temporary denial (Kaplan 2016, 57).

B. Operational Level

In contrast to the strategic level, there are examples of the operational use of cyber capabilities in conflict. Five cases come to mind: Syria 2007, Georgia 2008, Ukraine 2014–, Syria 2013– and one case of non-use of cyber capabilities in Libya 2011.

Operation Orchard (also known as Operation Outside the Box) took place in Syria in 2007, in which Israeli hackers disabled a Syrian anti-aircraft radar in Tell Abyad and then, in quick succession, launched a kinetic air-strike. The Israeli air force then destroyed a nuclear test site in Deir ez-Zor in northern Syria. The operations were successful and the digital component played a significant role in allowing Israeli F-15 jets to enter airspace unnoticed (Rid 2012, 19). Operation Orchard is an example of the sequential use of cyber capabilities as an "enabler" for kinetic operations, as well as a first-strike use. In such a case, the cyber operation produces an effect that is necessary for a subsequent kinetic operation.

The opposite is the joint or synchronous use of kinetic and cyber capabilities in the same context, where both components perform different functions. The Russian invasion of Georgia in 2008 and the conflict in eastern Ukraine (2014–present) are examples. The physical component of the Georgian conflict officially began on 7 August 2008 over a dispute in South Ossetia. Three weeks before this, the Georgian government and financial sector websites, along with various communication platforms, were hit by distributed denial of service (DDoS) attacks. This was a dress rehearsal for another wave of cyber attacks that were carried out simultaneously with the invasion of Russian combat troops. This time the goal was to impair Georgian communication with the outside world. Targets in the Georgian city of Gori, such as local news sites, were crippled by DDoS attacks just before Russian planes reached the city (Hollis 2011). In addition, an information operation component in the form of defacement of Georgian websites was used to spread chaos and uncertainty. Critical infrastructures, however, were not attacked. The complexity of these attacks can also be described as low. The Georgia incident demonstrates the lead time that cyber operations must have in order to be effective (Hollis 2011).

Integrating conventional and cyber operations to create joint effects is a challenge that many cyber powers are currently trying to figure out. A study by Nadiya Kostyuk and Yuri Zhukov, examining the use of cyber and kinetic military operations in Syria (2013) and Eastern Ukraine, shows that timing often does not work in sequential

or synchronous operations. Between 2014 and 2016, more than 1,841 cyber attacks and more than 26,289 kinetic operations were measured in Ukraine, but only a few of them occurred simultaneously. Instead of working together, physical and cyber operations took place largely separate from another, not creating joint effects. There was no reciprocity or strategic interaction between the two forms of attack. There was also no visible correlation between successful digital attacks from one side and kinetic counter-reactions from the other. This suggests massive synchronization problems and a low military shock effect (Kostyuk and Zhukov 2017). James Lewis argues that Russian cyber operations in Ukraine have failed to produce tactical or operational military effects beyond an initial tactical surprise effect (Lewis 2015). However, psychological effects, like sowing confusion and uncertainty, might be desired effects of cyber operations. Similar findings could be replicated in the Syrian conflict in 2013 (Kostyuk and Zhukov 2017). This suggests that operational cyber capabilities are (at the moment) an ineffective tool for exercising power in conflicts. However, if forces continue to train and exercise joint operations, this might change in the future. Coordination seems particularly challenging for states that rely on external proxy actors for cyber attacks, as was potentially the case in Ukraine.

To better understand the limitations of cyber operations on the operational level, it is worth looking at a case of non-use of cyber capabilities. Shortly before the start of the NATO operation to implement a no-fly zone in Libya in 2011 (Operation Odyssey Dawn), the US discussed the use of cyber operations but ultimately decided against it. The aim was to disable the Libyan air defense, which posed a threat to NATO aircraft. According to a New York Times report, the goal was similar to Operation Orchard: to disable or jam air defenses (Schmitt and Shanker 2011). The plan was rejected for several reasons. Firstly, the Obama administration feared that it would set a precedent that would have legitimized comparable actions by Russia and China. Second, the Americans did not have enough preparation time. This confirms the previously mentioned "cold-start problem" of cyber capabilities. The US Cyber Command did not have targets to strike or suitable malware for the relatively antiquated Libyan air defenses. Thirdly, it was uncertain whether such cyber attacks could have been carried out sustainably over a longer period. There were also doubts about whether cyber capabilities could reliably disable air defenses. There is also always a degree of uncertainty around whether a disabled system may recover more quickly than anticipated. Hence, large-scale cyber operations with a kinetic component, such as Stuxnet, have to be tested in simulated environments. This has implications for cyber warfare, where there is often no time for testing. If it is difficult to assess the impact of a cyber operation, military planners are hesitant to use it. If they have the alternative of destroying an asset permanently instead of using a potentially unreliable cyber capability, they tend to choose the former (Fink, Jordan, and Wells 2014). This is why the Libyan air defense system was permanently eliminated with cruise missiles.

Lastly, there was a desire not to waste highly complex and costly US cyber capabilities on the relatively low-tech Libyan forces and run the risk of their exposure (Schmitt and Shanker 2011). Cyber operations like Stuxnet have shown that there is a risk of losing assets because of malware spreading in an uncontrolled fashion.

C. Tactical Level

Not much is known about the tactical use of cyber capabilities; however, journalist Shane Harris has done extensive research on the use of offensive tactical cyber operations during counterinsurgency operations in Iraq in 2007 (Harris 2015). This operation had three components. First, the NSA correlated the phone metadata of Iraqi internet service providers with geographic maps and thus was able to pinpoint the geolocation of mobile phones used to trigger improvised explosive devices (IEDs). The NSA was able to destroy some of these from afar or to get the location of insurgents close by (Harris 2015, 69–72). This is an instance of tactical cyber as a counter-force capability.

The second component of the operation involved the use of malware against the insurgency's computer systems. Two variants were used here. The first involved the large-scale infection of numerous Iraqi users via manipulated phishing emails. The second involved the targeted infection of computers via USB sticks, which were carried by tactical cyber units in the field. The aim was to compromise the enemy information and communication or command and control network *Obelisk*, a kind of Al Quaeda Intranet (Harris 2015, 31).

The third component consisted of information operations against insurgents. With access to the Iraqi telephone network, US troops sent fake text messages to insurgents to demoralize them or to set a trap. For example, meetings were arranged where the person who appeared was captured. Malware was also used to locate individuals who uploaded propaganda videos via internet cafés (Harris 2015, 3–25).

Tactical cyber operations are subject to numerous restrictions, which explains why they have been used only sparsely. In most cyber nations, the use of offensive capabilities is decided at the strategic level, i.e. at a high point in the military chain of command. However, strategic cyber capabilities cannot simply be converted for tactical use at lower echelons in the chain of command because the use context is different (Metcalf and Barber 2014). Tactical cyber operations are difficult to integrate into the traditional target cycle of conventional forces due to their long planning and development time. Traditional weapons only need to be targeted once; tactical cyber operations must provide permanent covert access to a hacked system. However, this can be discovered by the defender, which can lead to a loss of access. Tactical cyber operations are therefore far more resource-intensive in their planning (Fink, Jordan,

and Wells 2014). The probability of discovering a hidden capability also influences their modality of deployment. It is pointless to invest large sums of money in a covert, tactical capability if it becomes uncovered in the first mission and thus becomes ineffective. Confidentiality requirements and tactical deployment have always been in conflict, as in combat situations, for example, the equipment can be captured by the adversary.

Unlike micro drones, mortars or anti-IED devices, tactical cyber capabilities are difficult to standardize, package and carry around. The tailoring requirement of cyber capabilities is a contradiction to the requirements of troops in the field. They need tools that must be repeatedly and reliably usable: an anti-IED device that only works against a certain type of mobile phone is less valuable than one that works against all types of mobile phones. Due to these characteristics of cyber capabilities, they are less suitable for tactical units (Porche et al. 2017, 47–50).

As with all cyber capabilities, collateral damage is difficult to anticipate. It is conceivable that tactical cyber operations in the field against computers of insurgents could also affect all other computers worldwide that have a similar configuration. In addition, civil infrastructure can be unintentionally affected, which can quickly become a PR disaster in tense foreign missions where the population is critical of foreign forces (Porche et al. 2017, 47–50). Tactical deployment can thus strategically escalate, for example, if collateral damage occurs worldwide. The general problem is that cyberspace does not match the geography of the battlefield on the ground. Conventional operations may be locally limited, but cyberspace is not (Metcalf and Barber 2014).

Lastly, lessons learned in Afghanistan and Iraq show that in difficult environments, such as vast landscapes and deserts, technology tends to fail. For tactical cyber operations to work, a data connection with enough bandwidth must exist. Computers need electricity and therefore they tend to be unreliable in combat situations, especially if the adversary possesses electronic warfare capabilities. Rebel forces with AK-74 rifles and almost no digital infrastructure still tend to be the most likely adversary in most asymmetric conflicts, and tactical cyber is limited against these common adversaries. For cyber operations in the field, certain proximity to the target is usually required. An enemy WLAN can only be hacked within the radio wave range. Tactical cyber operations in the field therefore only make sense if there is spatial proximity (urban warfare), if the desired effect can be standardized and thus made repeatable, if the required expertise is not too high, and if the effects can be limited to the local proximity.

4. DISCUSSION

The preliminary conclusion is that two major variables affect the utility of cyber technologies in war: the *timing* and operational *complexity* of cyber operations. Timing refers to questions of when and how long to engage in cyber operations to maximize effects. Operational complexity describes how hard it is to pull off the entire operation. Operational complexity includes various aspects such as the number of targets (one system vs. hundreds of systems to be hit at the same time), the defense level of the targets (multiple open attack surfaces vs. air-gapped systems), the availability of resources (intelligence and malware stockpile) as well as the size and internal organization and coordination of attacker teams.

Hypothesis 1: First-strike and sequential use of cyber capabilities seem easier to pull off, even for low-capacity actors, because the force-synchronization required for parallel use is hard to achieve.

In most of the analyzed cases, cyber attacks have been used in the early stages of a conflict. Cyber as a first-strike option in a conflict seems more promising and easier to pull off than continuous use in an ongoing conflict. Cyber attacks usually work best when they are not expected and when the adversary is unprepared. Continuous use requires a streamlined malware development cycle and enough personnel to rewrite malware after it gets burned or patched. If more malware gets burned than is reproduced, an operation is expected to slow down.

One aspect of operational complexity is the *availability of intelligence* that is needed to gain access to any hard-to-hit targets, especially military ones. The cases of non-use show that if there is no reliable intelligence on targets, cyber operations become riskier and less feasible. Intelligence collection and network reconnaissance involve an often time-consuming process, especially against highly secure, air-gapped targets, where in some cases, human intelligence is required. Even large cyber forces cannot prepare against any conceivable adversary, especially considering non-state actors and cyber proxies of which often little intelligence exists.

Hypothesis 2: The more preparation time there is, the more likely is the success of a cyber operation.

The case of Libya showed that if an attacker does not have time to tailor attacks for the specific targets, cyber operations are not feasible. Likewise, in rapidly unfolding crisis situations where there is no time to prepare and train, cyber tends to be of limited utility. Strategic cyber attacks aimed at shutting down an entire nation require large amounts of preparation time, as Nitro Zeus showed. But also, the cyber attacks against

Georgia had to be prepared and tested weeks in advance. How long it takes to prepare a cyber operation is also a function of the organization of one's cyber forces. Larger teams can probably produce greater malware stockpiles in a shorter amount of time and thus may need less preparation time compared to smaller teams. Larger teams, due to division of labor and functional differentiation, can also undertake multiple tasks or phases of an operation, such as reconnaissance and malware writing and testing, more efficiently, whereas smaller attack teams probably face some restrictions in the number of targets they can penetrate simultaneously or over a sustained period. Of course, this depends on their effectiveness and the structure of their organization. However, larger attack teams are potentially harder to synchronize than smaller teams. If states rely on external proxy actors like patriotic hackers, it may be harder to synchronize and control their attacks. The more actors are involved in a cyber operation, the higher the complexity becomes.

Hypothesis 3: High operational complexity increases the risk of failure of any sustained cyber campaign.

Coordination of two military components, such as a cyber force and an air force, in one single operation against one target, like Operation Orchard, seems manageable. The more military components or organizations that come into the loop, the harder it becomes to coordinate them. The more actors are involved and the longer an operation lasts, the more complex it tends to get. The broader the scope of the operation, i.e. striking a single target vs. striking an entire nation over a period of time, the more complex the operation. The same is true for targets with broader attack surfaces. As many IT-systems are interdependent, there is always a risk of unexpected collateral damage when shutting these down with cyber attacks. As in any complex system where the interaction of the different individual parts is non-linear and opaque, it is hard for external observers to make predictions. Thus, the more complex cyber operations get, the harder it becomes to predict outcomes, and thus the higher the uncertainty and the lower the ability to guarantee success.

Hypothesis 4: If military commanders have alternative options to cyber operations with high complexity and thus uncertain reliability, they tend to choose the safer option (that is, using kinetic means to disable targets instead).

The high degree of uncertainty of complex cyber operations also influences the use decision of commanders. Libya and Nitro Zeus showed these signs of hesitation. Since the damage of cyber attacks is often temporary, there is always a risk of unanticipated resilience. A shut-down system can come back online quicker than anticipated. However, if a cyber attack is the first step in a whole military war plan and this step fails, the rest of the planning that depends on the effects of the first cyber attack is

at risk. Therefore, traditional means of physically destroying targets may seem more reliable.

These hypotheses will be tested in future research. The preliminary conclusion is that the argument of the all-purpose sword does not hold up completely. Cyber technologies in war certainly have some benefits, but a lot of operational hurdles need to be overcome for them to become a perfect all-purpose sword. Right now, it seems that cyber operations are more like a specialized weapon for quick strikes, rather than for lengthy and sustained campaigns. They require a lot of training and preparation and are difficult to wield together with another type of arms. As with all weapon types, in the end, the organizational structure and the tactics used are what determines the success rate of any given weapon.

REFERENCES

Ablon, Lillian, and Andy Bogart. 2017. *Zero Days, Thousands of Nights: The Life and Times of Zero-Day Vulnerabilities and Their Exploits*. Research report RR-1751-RC. Santa Monica, Calif: RAND. https://www.rand.org/pubs/research_reports/RR1751.html.

Arquilla, Jon. 2017. "The Rise of Strategic Cyberwar?" https://cacm.acm.org/blogs/blog-cacm/221308-the-rise-of-strategic-cyberwar/fulltext.

Bateman, Robert. 2015. "Understanding Military Strategy and the Four Levels of War: When 'Strategy' Gets Thrown Around by Politicians and the Media, You Can Bet It's Being Misused." https://www.esquire.com/news-politics/politics/news/a39985/four-levels-of-war/.

Borghard, Erica D., and Shawn W. Lonergan. 2017. "The Logic of Coercion in Cyberspace." *Security Studies* 26, no. 13: 452–81. https://doi.org/10.1080/09636412.2017.1306396.

Clausewitz, Carl von. 1982. *On War*. Reissued. Penguin Classics. London: Penguin Books.

Crane, Alfred C., and Richard Peeke. 2016. "Using the Internet of Things to Gain and Maintain Situational Awareness in Dense Urban Environments and Mega Cities." *Small Wars Journal* (February).

Defense Industry Daily. 2020. "Tomahawk's Chops: XGM-109 Block IV Cruise Missiles." https://www.defenseindustrydaily.com/block-iv-xgm-109-tomahawk-chopped-07423/.

Fink, Kallie D., John Jordan, and James E. Wells. 2014. "Considerations for Offensive Cyberspace Operations." *Military Review* (May-June).

Gartzke, Erik. 2013. "The Myth of Cyberwar: Bringing War in Cyberspace Back down to Earth." *International Security* 38, no. 2: 41–73. https://doi.org/10.1162/ISEC_a_00136.

Harris, Shane. 2015. *@War: The Rise of the Military-Internet Complex*. First Mariner Books edition. Boston, New York: Mariner Books Houghton Mifflin Harcourt.

Hollis, David. 2011. "Cyberwar Case Study: Georgia 2008." *Small Wars Journal*. https://smallwarsjournal.com/jrnl/art/cyberwar-case-study-georgia-2008.

Jun, Jenny, Scott LaFoy, and Ethan Sohn. 2015. *North Korea's Cyber Operations: Strategy and Responses*. Washington, DC, Lanham, Boulder, New York, London: Center for Strategic & International Studies; Rowman & Littlefield.

Kaplan, Fred M. 2016. *Dark Territory*. New York: Simon & Schuster Paperbacks.

Kostyuk, Nadiya, and Yuri M. Zhukov. 2017. "Invisible Digital Front: Can Cyber Attacks Shape Battlefield Events?" *Journal of Conflict Resolution* 63, no. 2: 317–47. https://doi.org/10.1177/0022002717737138.

Lawson, Sean. 2013. "Beyond Cyber-Doom: Assessing the Limits of Hypothetical Scenarios in the Framing of Cyber-Threats." *Journal of Information Technology & Politics* 10, no. 1: 86–103. https://doi.org/10.1080/19331681.2012.759059.

Levy, Jack S. 2008. "Case Studies: Types, Designs, and Logics of Inference." *Conflict Management and Peace Science* 25, no. 1: 1–18. https://doi.org/10.1080/07388940701860318.

Lewis, James Andrew. 2015. "'Compelling Opponents to Our Will': The Role of Cyber Warfare in Ukraine." In *Cyber War in Perspective: Russian Aggression Against Ukraine*, edited by Kenneth Geers, 39–48. Tallinn, NATO CCDCOE.

Libicki, Martin C. 2009. *Cyberdeterrence and Cyberwar*. Santa Monica: RAND Corporation. http://swb.eblib.com/patron/FullRecord.aspx?p=566752.

Metcalf, Andrew, and Christopher Barber. 2014. "Tactical Cyber: How to Move Forward?" https://smallwarsjournal.com/jrnl/art/tactical-cyber-how-to-move-forward.

Pape, Robert Anthony. 1996. *Bombing to Win: Air Power and Coercion in War*. 1. publ. Cornell Studies in Political Economy. Ithaca, NY: Cornell University Press.

Porche, Isaac, Christopher Paul, Chad C. Serena, Colin P. Clarke, Erin-Elizabeth Johnson, and Drew Herrick. 2017. *Tactical Cyber: Building a Strategy for Cyber Support to Corps and Below*. Research report RR-1600-A. Santa Monica Calif. RAND.

Rid, Thomas. 2012. "Cyber War Will Not Take Place." *Journal of Strategic Studies* 35, no. 1: 5–32. https://doi.org/10.1080/01402390.2011.608939.

Sanger, David. 2018. *The Perfect Weapon: War, Sabotage, and Fear in the Cyber Age*. New York, Melbourne, London: Crown Publishers; Scribe.

Sanger, David, and Mark Mazzetti. 2016. "U.S. Had Cyberattack Plan If Iran Nuclear Dispute Led to Conflict." *New York Times*, February 16. https://www.nytimes.com/2016/02/17/world/middleeast/us-had-cyberattack-planned-if-iran-nuclear-negotiations-failed.html.

Schmitt, Eric, and Thom Shanker. 2011. "U.S. Debated Cyberwarfare in Attack Plan on Libya." *New York Times*, October 17. https://www.nytimes.com/2011/10/18/world/africa/cyber-warfare-against-libya-was-debated-by-us.html.

Smeets, Max. 2018a. "A Matter of Time: On the Transitory Nature of Cyberweapons." *Journal of Strategic Studies* 41, no. 1-2: 6–32. https://doi.org/10.1080/01402390.2017.1288107.

Smeets, Max. 2018b. "The Strategic Promise of Offensive Cyber Operations." *Strategic Studies Quarterly* (Fall): 90–113.

US Air Force. 2019. "Annex 3-70-Stragegic Attack." https://www.doctrine.af.mil/Doctrine-Annexes/Annex-3-70-Strategic-Attack/.

Valeriano, Brandon, Benjamin M. Jensen, and Ryan C. Maness. 2018. *Cyber Strategy: The Evolving Character of Power and Coercion*. New York, NY: Oxford University Press.

Valeriano, Brandon, and Ryan C. Maness. 2015. *Cyber War Versus Cyber Realities: Cyber Conflict in the International System*. New York, NY: Oxford University Press.

van Haaster, Jelle. 2019. *On Cyber: The Utility of Military Cyber Operations During Armed Conflict*, Amsterdam, University of Amsterdam.

Young Kong, Ji, Kyoung Gon Kim, and Jong in Lim. 2019. "The All-Purpose Sword: North Korea's Cyber Operations and Strategies." In *11th International Conference on Cyber Conflict: Silent Battle*, edited by T. Minárik, S. Alatalu, S. Biondi, M. Signoretti, I. Tolga, and G. Visky, 143–62. Tallinn: NATO CCD COE Publications.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Correlations Between Cyberspace Attacks and Kinetic Attacks

Martin C. Libicki
Distinguished Visiting Professor
Center for Cyber Security Studies
U.S. Naval Academy
Annapolis MD, USA
libicki@usna.edu

Abstract: Although confrontations in cyberspace can conceivably stay in cyberspace (or at least not involve violent conflict), they can also become entangled with confrontations in the physical world. This paper explores how, by raising the following questions: (1) Do countries retaliate in the real world for operations in cyberspace? (2) Would countries make an equivalence between the damage from cyberattacks and from physical attacks (in ways that could spill over from the one to the other)? (3) Does cyberspace escalation lead to kinetic escalation and is the reverse also true? (4) Can cyberspace operations against sensitive targets put them in play for kinetic operations? (5) Would the failure to react to cyberattacks embolden attackers to carry out kinetic attacks? This paper leverages what is known and what can be logically assumed about cyber operations, notably by drawing lessons from Russia's use of cyberspace operations in Georgia and Ukraine, Iran's cyber and physical attacks against Saudi Aramco, and China's military doctrine vis-à-vis U.S. space assets. The broad conclusion is that, so far, conflict in cyberspace rarely echoes into the world of kinetic conflict (although kinetic conflict increasingly has cyberspace dimensions). This raises the question of whether and why a threshold is emerging between non-lethal and lethal attacks.

Keywords: *escalation, cyberattack, kinetic*

1. INTRODUCTION

Although interstate confrontations in cyberspace *could* stay in cyberspace (or at least within the information domain), nothing mandates that both sides will observe such boundaries. Cyberattacks can become entangled with more conventional military operations, as they have in Georgia (2008) and Ukraine (2014-).

If this is true, *the prospect that conflict in cyberspace can bleed over into kinetic conflict suggests that operations in cyberspace have the potential to cause more serious instability than assumed* (e.g. Healey 2019). But, is it true?

To explore the issue, we look at the relationship between incidents and escalation in cyberspace and their counterparts in the physical world by posing five sub-questions:

- Do countries retaliate in the real world for operations in cyberspace?
- Would countries make an equivalence between the damage from cyberattacks and from physical attacks (in ways that could spill over from the one to the other)?
- Does cyberspace escalation lead to kinetic escalation and is the reverse also true?
- Can cyberspace operations against sensitive targets put them in play for kinetic operations?
- Would the failure to react to cyberattacks embolden attackers to carry out kinetic attacks?

We will try to use (known) past events to address these questions. That said, there are not many incidents to work with. Although scholars have compiled large datasets of cyberspace incidents (see, in particular, Valeriano, Jensen, and Maness 2008), the bulk of them, by far, are acts of cyberespionage, and many of the rest are Distrubuted Denial of Service (DDOS) attacks. The paucity of examples relevant to escalation necessarily limits how robust any answers are to future events.

2. DISTINCTIONS AND CAVEATS

Despite the prominence of "persistence" as reflected in the phrases "advanced persistent threat" or "persistent engagement", conflict in cyberspace tends to have an episodic quality. There is no good equivalent to holding or contesting land (persistent access to a system is only slightly analogous). Incidents of cyberspace conflict are often *sub rosa*, and usually unacknowledged. They have not followed declarations of war or any of its modern equivalents. As a practical matter, it is hard to judge whether

any one cyberspace operation – especially one in a long series of similar events – is or is not escalatory.

Escalation, itself, has been defined as "an increase in the intensity or scope of conflict that crosses threshold(s) considered significant by one or more of the participants" (Morgan et al. 2008). This formulation contains two key elements.

One element is the ability to measure the *intensity* of cyberspace operations. This requires, at a minimum, two such operations of a similar type between the same combatants and in the same or similar context – plus some metric that indicates that one is more serious than the other. But intensity is a measure of effort, not success. It has to be inferred from a set of incidents whose effects reflect not only intensity but other factors such as the quality of defense. In other words, while one side may have increased the intensity of its efforts, the other side may not perceive as much if its defenses have risen to the challenge. Perhaps the best that can be said of a cyberattack is that it may be considered akin to escalatory if it is unexpected or at least unprecedented in a particular context.

A second element in the definition is the existence of significant thresholds. It is unclear whether there are *any* such thresholds in cyberspace, *per se*. There is no broad consensus as to what targets are off limits. In 2015, the UN Group of Government Experts tried to put civilian infrastructures off limits (United Nations 2015), but power grids *have* been attacked by at least one great power since then (Goodin 2015 and 2017). Putatively, there may also be a recognized threshold between a cyberattack with casualties and one without. But no cyberattack has caused direct casualties and determining indirect casualties is difficult. Indirect casualties are subject to dispute: e.g., do two reported suicides after exposing the customers of the Ashley Madison website count (Baraniuk 2015)? Even if Wannacry was associated with higher-than-expected death rates in U.S. hospitals, the details are hidden in litigation and a close review of death rates in Britain's National Health Service shows no discernable net effect (Ghafur et al. 2019). Although the definition of escalation may be fulfilled if one side (typically, the target) believes that a cyberattack has crossed the line, countries have been slow to determine or at least announce what such lines might be. Calling a cyberattack escalatory if it produces unexpected (and, generally, more severe) consequences may be close enough to right.

A third element is the challenge of determining whether one attack was a response to another. This is particularly difficult in cyberspace, where a cyberattack carried out as a response may require establishing accesses in a target system, a process of hard-to-predict length. In some cases, it may require establishing a capability; Iran could

not respond to Stuxnet in 2010 until it had built capabilities that it deployed in 2012. However, generating a *kinetic* response to a cyberattack would seem to take less time.

One last caveat. In physical combat there is a *rough* correspondence that relates effort to effects and effects to perceptions (of effects) – despite the fog and friction of war or the difficulties of battle damage assessment. Time creates the opportunities to sort out much of what initially appears ambiguous. In cyberspace, most probes fail, many go unnoticed, and even successes are not always immediately discovered. Stuxnet, for example, which clearly destroyed Iranian centrifuges, was not detected as a cyberattack until the summer of 2010 even though its effects took place in late 2009 and early 2010. Only later did Iranians come to understand that the failures they were definitely seeing resulted from deliberately corrupted commands rather than accidents, poor operational procedures, or substandard components. Malware implants present a particular problem. Finding one in a target system may offer little indication of what its purpose was: e.g., espionage or cyberattack? If the purpose is obscure, the intent will be at least as obscure. This leaves in doubt whether the other side intended to escalate. And even a more fully-completed cyberattack which shows that, say, a system's controls can be usurped, does not prove whether the point was to test procedures, brandish capabilities, or wreak damage. Arguably, the late 2016 Russian cyberattack on the Ukrainian power grid was deliberately stopped once the point was made, when it could have gone far longer (Greenberg 2019). Again, we can only work with what we have. This may explain why the topic can use further exploration despite good conceptual work having been done (Borghard and Lonergan 2017; Lin 2012).

3. PROPOSITION: CYBERATTACKS CAN LEAD TO KINETIC RETALIATION

Were this true, then a sufficiently grave cyberattack could have serious escalatory consequences by crossing the boundary into what is commonly recognized as armed conflict. In 2009, an anonymous U.S. administration source asserted that "If you shut down our power grid, maybe we will put a missile down one of your smokestacks" (Gorman and Barnes 2011). Accordingly, a cyberattack would beget kinetic retaliation, which begets more kinetic retaliation, which evolves into a war, and, if at least one side has nuclear weapons, a chance, albeit very small, of nuclear Armageddon.

Nothing *so far* suggests this as a plausible scenario. True, small-scale tit-for-tats in cyberspace (mostly web defacements and DDOS attacks) have taken place between the usual dyads (e.g., India and Pakistan or Israel and Palestinians). But the only significant retaliation for a cyberattack – and not everyone sees it that way – has been

the Iranian DDOS attacks on U.S. banks in late 2012 as a response to the Stuxnet attack of 2010 and (with somewhat less clarity) Iran's use of wiper malware against Saudi Aramco and Qatar after a similar wiper attack on its refineries (Zetter 2015).

The shift from cyberattacks to violent attacks has so far been scarce and ambiguous. One *possible* incident was the violent death of Mojtaba Ahmadi, the commander of Iran's Cyber War Headquarters, several weeks after the traffic controls of Haifa's Carmel Tunnel had been hacked (see InfoSecurity 2013; McElroy and Vahdat 2013). Even though the cyberattack preceded the death, Israel's announcement came afterwards. Both Israel and Iran deny the connection, however. Another was a physical attack on a Gaza building said to house Hamas hackers – but such claims could be dismissed as an opportunistic justification of a particular bombing attack, carried out during a conflict in which bombing attacks of all sorts were frequent (Chesney 2019).

Conclusion: a kinetic retaliation to a cyberattack is possible but cannot yet be deemed a likely consequence.

4. PROPOSITION: CYBERATTACKS WILL BE TAKEN AS SERIOUSLY AS EQUALLY DAMAGING KINETIC ATTACKS

If countries react to cyberattacks as they would to equivalent kinetic attacks, then an escalation in cyberspace (defined as above) could well result in a comparable escalation in physical space – again with the expected effects on international stability.

But would they? Much of the answer depends on what constitutes "comparable." Kinetic military effects tend to include death and destruction. No cyberattack has killed anyone directly, and few have actually broken physical things; wiping a hard drive – as many cyberattacks have done – still leaves the hard drive physically intact. But cyberattacks have been quite costly to their victims, even if measured solely in disruption and remediation costs. The NotPetya attacks were said to have cost their (mostly corporate) victims up to $8 billion (Greenberg 2018). Putatively, a kinetic attack that destroys $8 billion worth of military equipment but harms no one would be comparable and should, one would imagine, bring about a comparable reaction.

Imagining a kinetic attack that breaks things but hurts no one used to be an exercise in fantasy. But the Iranian take-down of a $150 million U.S. Global Hawk in the summer of 2019 *was* such an attack. The U.S. response, a cyberattack, was also non-lethal. Lest this choice of avoiding lethality be ascribed to the individual characteristics of the U.S. President, note that the Pentagon was also thinking along similar lines. One

of its favored options was to sink an Iranian craft, but only *after* giving its sailors time to get away (Baker, Schmitt, and Crowley 2019). Earlier, a Turkish shootdown of a Russian jet near the Syrian border had drawn a cyber response (e.g., DDOS attacks), but nothing violent (Murgia 2015).

Returning to NotPetya, the U.S. reaction to this costly event was a limited set of sanctions. If Russia had deliberately disabled commercial satellites whose total replacement value summed to that much, would the United States have also limited its response to sanctions? One might counter that many of the affected corporations were not U.S.-headquartered: for example, Maersk, a Danish shipping company. If that matters, then replace United States by NATO and re-ask the question. So, while the non-lethality of cyberattacks means that a plausible response would be non-lethal, the failure to respond to NotPetya suggests that the broad scope of the cyberattack may have also played a role. Perhaps a cyberattack that damages software and thereby levies costs on victims is different in kind from a comparably costly kinetic attack that damages hardware.

Research by Professor Jacqueline Schneider casts further doubt that a cyberattack would be treated as tantamount to a comparable kinetic attack (see Kreps and Schneider 2019). The results of two exercises – one conducted at the U.S. Naval War College and the other on-line – suggest that cyberattacks introduced into a simulated crisis were more often ignored or, at most, motivated a weak response in comparison to comparable kinetic attacks.

In fairness, the United States has been used as the exemplar of how countries may respond to cyberattacks and other countries may react differently. But the United States deserves attention because it has responded most overtly, whether through public statements, levied sanctions, or news reports (Israel is also active, but it is a far smaller country and unique in many relevant respects). It is unclear whether the difference is that the United States suffers more cyberattacks than other countries (or seems to in part because of uncensored media coverage) or whether other countries have covert ways of responding that are not widely known. That noted, Jensen and Valeriano (2019) indicate that when citizens of the United States, Russia, and Israel were given a scenario with a major cyberattack, only a small percentage chose to escalate as a result. Roughly half of the respondents wanted something less than a tit-for-tat response. They concluded that, "to date, cyber operations have tended to offer great powers escalatory offramps".

Conclusion: cyberattacks would be deemed less likely to garner a kinetic response than would kinetic attacks that levy comparable costs, because they are generally non-lethal and somehow considered less serious and more easily recovered from.

5. CYBERATTACKS PRESAGE KINETIC ATTACKS

An opening attack by a country that is adept at cyberattacks against a country that depends on information systems could be a precursor to a broader armed attack. Cyberattacks, especially against a surprised – hence unprepared – target, have some potential to blind, confuse, and even disarm the adversary, making conventional victory easier. Cyberspace theorists from James Mulvenon onward have posited a Chinese military campaign whose first move is to paralyze the U.S. ability to move warfighters and materiel across the Pacific, giving China additional time to take and consolidate military objectives in East Asia before the United States arrives in force. One advantage of using cyberattacks this way is that the ambiguity about their characteristics (while the target asks: why are systems failing?) and their attribution can retard the target's conclusion that it must prepare for immediate war. By contrast, a kinetic attack (e.g., against sensors) initiated as a prelude to wider hostilities would more certainly remove the element of surprise when the wider hostilities commenced.

The best case that cyberattacks do precede kinetic attacks comes from the Russia-Georgia war in 2008. Just prior to the onset of that conflict (dating from when Russian troops moved into Georgia and not into South-Ossetia, a part of Georgia outside its government's *de facto* control), DDOS attacks from Russian sources (probably but not provably state-directed) limited Georgia's access to the Web, notably preventing the government from putting out its view of the conflict. Russian cyber or at least electronic interference may have deliberately hindered Georgia's mobile phone system, which had military uses. By contrast, Russian DDOS attacks on Estonia (which may or may not have been state-directed) *followed* the first night of riots by ethnic Russians in Tallinn (April 26, 2007) which themselves were prompted by Estonia's decision to move the "Bronze Soldier" from downtown to a nearby military cemetery. And these DDOS attacks did not precede any kinetic military operations by Russia against Estonia.

Cyberattacks – with the important exception of DDOS attacks – typically require months of planning. Unpredicted kinetic conflicts are thus unlikely to be preceded by cyberattacks. In 2011, NATO aircraft were engaged over Libya, and the threat that they would be shot down by Libyan surface-to-air missiles (SAMs) reportedly prompted discussion of using cyberattacks to neutralize these SAMs (Nakashima 2011). Ultimately, no such cyberattack took place (as far as publicly revealed). By the time it could have been completed, there would have been little need to suppress Libyan SAMs. If NATO had anticipated in advance having to fight in Libya – riots in Tunisia that set off the Arab Spring did precede NATO operations over Libya by three months – it might have had time to disable Libyan SAMs by cyberattack, but hindsight never needs glasses. Conversely, if cyberattacks *are* used to precede kinetic

combat, or even in the first days of kinetic combat – and their effects are detected and correctly attributed – then it would be difficult for the cyberattacker to claim that war had been a complete surprise to it. The timing of events would suggest that the cyberattacker had assessed the possibility of war as being likely enough to justify laying in cyberattack preparations. This would strain any argument that the target (of the cyberattacks) had started the war out of the blue.

In a putative future in which every major country has implants in the military systems of anyone they have the remotest chance of having to fight against, then a cyberattack may not be so indicative. But that has (probably) not yet happened. What is likely to come first is that major countries will be distributing implants into adversary networks for cyberespionage – which, after all, is a normalized peacetime activity that friends do even to friends. But though, in theory, every cyberespionage penetration is a potential cyberattack penetration, the immediate targets will be different. Because the knowledge possessed by SAM systems has limited intelligence value, such systems are rarely first-tier targets for cyberespionage. As a rule, the knowledge necessary to cause specific types of failures (e.g., how to make a centrifuge spin itself to death) must be acquired specifically for that purpose. If reports are reliable, however, the United States *has* placed implants in the military systems of countries it may have to fight. The aforementioned Iranian shootdown of a Global Hawk missile was, ultimately, followed by a U.S. cyberattack on Iran's ship-tracking database. In all likelihood, the path to the database was laid in *before* Iran shot down the Global Hawk (mid-June 2019) and may have been laid in even before Iran (re)started targeting commercial shipping (mid-May 2019) – although if Iran's networks were easy to penetrate, preparatory intrusions could have started not much earlier than the cyberattacks did. Before the Joint Comprehensive Plan of Action (JCPOA) agreement of 2015 with Iran, there were stories that the United States had laid in attacks against Iranian electrical infrastructures named *Nitro Zeus* (Sanger and Mazzetti 2016).

In the conflict between Russia and Ukraine, most of the major cyberattacks have come from Russia. Because the war would not have started but for the unexpected resignation of Ukraine's President Viktor Yanukovych, Russia did not accompany its kinetic operations with cyberattacks that required great planning – although it did carry out DDOS attacks and acts of electronic warfare from its outset. Over time, Russians did attack Ukraine's infrastructure and launched a notable supply chain attack (from whence NotPetya) against the Ukrainian company, MeDoc. But the attack on Ukraine's power grid did not take place until the second year of conflict. Detailed study of the Ukraine and Syria conflicts suggests that "cyber activities failed to compel discernible changes in battlefield behavior. Indeed, hackers on both sides have had difficulty responding to battlefield events, much less shaping them" (Kostyuk and Zhukov 2017).

There are no known examples, however, of the *target* of a surprise kinetic attack having pre-empted such an attack using cyberattacks.

Conclusion: although most cyberattacks do not presage kinetic combat, some cyberattacks might. Surprise cyberattacks by a cyberspace-adept country against a cyberspace-dependent country would offer the best opportunity for their usage, but many kinetic wars come as a surprise to both sides.

6. PROPOSITION: CYBERATTACKS MAY PUT HITHERTO SACROSANCT TARGETS IN PLAY FOR KINETIC ATTACKS

In WWII, cities were initially considered sacrosanct. Then Germany bombed Warsaw and later Rotterdam, but these targets could be considered of direct relevance to military operations on the ground. Then Germany bombed residential districts of London while allegedly going after air defense sites. Then both sides practiced unrestricted air warfare. Today, there is a broad, but not necessarily realistic, expectation that space systems and nuclear command-and-control systems are sacrosanct. One can easily imagine a conflict, perhaps one of local relevance only, in which such systems are initially considered off limits by both sides – only to be placed in-bounds by subsequent escalation. Such escalation may have many sources, but one potential source is that cyberattacks on space and/or nuclear command-control-and-communications (NC3) systems may put targets in play for kinetic attacks as well.

Space systems and NC3 systems really ought not to be accessible to cyberattack. Their military criticality should make anyone think twice about connecting them to the outside world, and they do not need the Internet to function. But these circumstances hardly provide proof against mischief – for the usual reasons. Not everyone understands how the fact of access alone heightens cybersecurity threats. The pressure to expand access to sensitive systems is often hard to resist, especially when expanding access can facilitate their support and maintenance. Not every access point is easy for defenders to discover; some system components have been given unadvertised connectivity at the factory or in the course of repair. People put great trust in protections (e.g., firewalls) that can be manipulated. So, while we lack documented evidence that any hacker has breached NC3 systems, it is too early to say for sure that they cannot be hacked (Futter 2018). And while cyberattacks on *military* space systems have not been reported, probable hostile penetration of the control systems of civilian satellites has been (see Barrett 2019; Leavitt 2011; Newman 2018; Tucker 2019).

To be fair, it is unclear how sacrosanct space systems really are from *kinetic* attack anyway. The United States, China, Russia, and India have all tested anti-satellite systems. And while all four have paid respect to the notion of peace in the heavens, none has foresworn being the first to use their anti-satellite systems. Finally, satellites can be destroyed without creating casualties, an argument in favor of their being targeted if military need arises.

The inviolability of NC3 systems in scenarios short of nuclear war is based on the proposition that nuclear stability requires the major powers to be assured of their second-strike retaliatory capability. Some (Acton 2018) fear that cyberattacks on systems that support command-and-control for both conventional and nuclear systems will seem motivated to reduce the target's nuclear retaliatory capability in the guise of legitimate warfighting. Such suspicions could lead, at best, to twitchy adversaries apt to overreact to any further threat to their capability – and, at worst, to adversaries concluding that they must use their nuclear weapons before they otherwise lose them. Accordingly, others (Danzig 2014) have proposed that the major powers pledge not to carry out cyberattacks on adversary NC3 systems – a proposition that, suffice it to merely note, is both laudable and problematic: enforcement would be difficult and might require banning NC3-directed cyberespionage, some types of which could provide reassurance against surprise attack.

But will cyberattacks on space or NC3 systems put them in play for kinetic attacks among the immediate combatants – or worse, create a precedent that colors how every other country might treat its foes' systems? Maybe not – in part because of the ambiguity and the non-lethal nature of cyberattacks. Ambiguity affects three questions. *One*, were the perceived effects the result of adversary action – in contrast to misunderstandings (e.g., of how the supposedly targeted system was supposed to work), design flaws, accidents, Mother Nature, etc.)? *Two*, if adversary actions were the cause, were they deliberate, or inadvertent (e.g., because of hacker mistakes, malware drift, etc.)? *Three*, was the cyberattacker the same adversary that is attacking in physical space? The importance of determining why systems failed is clear enough. Whether or not system failure was *deliberate* speaks to the other side's intent, whether it was sending a signal, and whether a repeat performance can be expected. The importance and the difficulty of attribution is clear enough, as well. Time also plays a role. With the Pearl Harbor attack, to give an example, its fact, its deliberateness, and its perpetrator were instantly obvious. Characterization, intentionality, and attribution in cyberspace may also be instantly obvious in some cases, but in other cases could take time to discover. Reaching a conclusion that action is merited may precede acquiring 100% confidence in that conclusion – and it may take a great deal of confidence to escalate a conflict when there is some lingering doubt that any such escalation was

forced on the target. In the, say, months in-between, matters may escalate for other reasons – or the conflict could end.

The non-lethal and often temporary nature of cyberattacks may also put off comparable escalation into unleashing kinetic attacks on satellites. All possible satellite attack modes are non-lethal (with two minor exceptions: attacks on the International Space Station and attacks whose debris causes ground casualties). And many physical attacks on satellite services – such as jamming, dazzling, or blinding – are temporary and, as such, likely to be carried out before contemplating escalating to destructive attacks. By contrast, many attacks that destroy satellites can endanger all other satellites by create long-orbiting space debris. So, the best guess at this point is that cyberattacks on satellites with reversible effects would be treated like temporary physical attacks. Cyberattacks that disable satellites permanently (e.g., by directing them to an unsustainably low orbit) – and such permanence may take a while to ascertain – would be considered more serious but not as serious as physical destruction. But that does not mean that cyberattacks would be shrugged off.

NC3 systems include a wide variety of components. Some are unmanned: e.g., satellites (for communications, and early warning), radar dishes, communications lines, transmission towers. Others are manned: e.g., radar stations, command centers, command authorities (e.g., key nuclear commanders wherever they are). Many can be disabled by cyberattacks, but few can be disabled permanently that way. The range of temporary attacks on NC3 based on physical effects (e.g., as dazzling, jamming, or blinding are for satellites) is limited compared to the array available against satellites. Finally, kinetic attacks against many NC3 components risk casualties. Conversely, the sacrosanct nature of NC3 systems is better established than with satellites. Thus, a best guess at this point is that cyberattacks against NC3 systems – provided they are confidently characterized and attributed – can open the door for kinetic attacks against NC3 systems. A lot will depend on whether the two parties are fighting a kinetic war at the time. If so, a response may come faster. If not, a lot may depend on events that intervene before a next kinetic war starts.

Conclusion: cyberattacks have the potential to put privileged assets in play for kinetic attacks, but not necessarily.

7. PROPOSITION: FAILURES TO RESPOND TO CYBERATTACKS EMBOLDEN KINETIC ATTACKS

A failure to respond could mean (1) doing nothing, (2) doing something that falls short of signaling seriousness (e.g., imposing individual economic sanctions after NotPetya), or (3) doing something that should impress the attacker but does not. Here, we focus on the possibility of an unanswered escalation in cyberspace emboldening escalation in physical space by the same actor.

The evidence here is mixed.

In the ongoing undeclared conflict between Saudi Arabia and Iran, cyberattacks on Saudi Aramco in 2012 wiped the memories of roughly 30,000 computers. The same attack may have tried to ruin physical (oil field) equipment but never reached that far (Perlroth 2012). Putatively, this may have been retaliation for a less-well-reported cyberattack on Iran's main oil export terminal (Reuters 2012). Neither Saudi Arabia nor the United States retaliated (as far as known). In the summer of 2019, Saudi Aramco facilities were hit by missile attacks; by the end of 2019 there had been no kinetic response nor any other response (also, as far as known). Did the lack of response to the 2012 cyberattack therefore encourage Iran to think it could get away with a physical attack? Note that in both cases, attribution was not instant. The 2012 cyberattack initially looked as if it could have been an inside job, until the consensus formed that it was Iran's doing. The 2019 missile attack on Aramco refining facilities was initially ascribed to Yemen's Houthi rebels, although later analysis indicates that the discerned direction of the incoming missiles was unlikely if launched from Yemen and that the Houthi rebels anyway lacked the technological sophistication to aim their weapons so accurately. Indirectly, and perhaps even directly, it was Iran's doing. Although the failure to push back hard on Iran's earlier kinetic attacks on neutral shipping and the U.S. Global Hawk may have persuaded Iran it could have gotten away with the missile attack, the failure to see much response from their 2012 cyberattack may have played a role in such assurance as well.

The Russo-Ukrainian conflict featured multiple cyberattacks. The NotPetya malware was designed to undermine trust in the products of Ukrainian corporations. Electrical systems were hacked twice. However, there have been essentially no *kinetic* attacks on Ukrainian infrastructure (at remove from the front lines in eastern Ukraine). Thus, the general failure of the West to respond to Russian cyberattacks on infrastructure does not seem to have encouraged Russia to launch kinetic attacks.

Conclusion: there is scant evidence *so far* that a failure to respond to cyberattacks, especially on critical infrastructure, puts them in play for kinetic attacks.

8. OVERALL CONCLUSIONS

Overall, there is little public evidence that hostile events in cyberspace echo strongly outside it. Indeed, rarely do events in cyberspace – much less escalation in cyberspace – lead to serious responses at all. Some research suggests that even severe cyberattacks would generally be less likely than kinetic attacks to induce a response. Although opening cyberattacks can precede kinetic attacks, there are also cases when war comes as a surprise and cyberattacks are not used until the proper accesses to target systems have been gained. Cyberattacks have the potential to put hitherto sacrosanct targets – notably space systems, and other NC3 elements – in play, but cyberattacks have reportedly taken place against satellites while kinetic attacks (weapons tests aside) have not, so far. The failure to respond to cyberattacks *may* have played a role in enabling missile attacks on Saudi Aramco facilities, but the link is distant (seven years earlier) and tenuous. There is no analog (yet) in the Russo-Ukrainian conflict.

Several reasons could be adduced to explain the lack of correlation. One is that while there *could be* cyberattacks consequential enough to induce echoes in the physical world, none have reached that threshold and it may well be that none *could* reach that threshold. Even as the attack surface for cyberspace operations keeps growing, hackers grow more talented, and their leaders more aware of the gains available from such operations – defense is not sleeping. Those who own networks are taking cybersecurity seriously (at long last), cloud computing may have helped put defenses in the hands of those for whom protection is a profit center, and the cybersecurity industry itself is robust. Succeeding generations of software – e.g., versions of Windows operation systems – are also more impervious to intrusions. Two is that, in common with many widely-feared phenomena, cyberattacks have evolved from an acute problem (one both rare and fearsome) to a chronic problem (more common, but something that one can adjust to). Three, the oft-expressed belief that cyberwar is war has yet to take hold. Because cyberspace operations are ambiguous (and not easily grasped even when clear) and their effects almost always temporary and not (yet) lethal, they may be considered something separate and apart. Time will tell whether this distinction will continue to be observed.

REFERENCES

Acton, James. 2018. "Escalation through Entanglement." *International Security* 43, no. 1 (Summer): 56–99.

Baker, Peter, Eric Schmitt, and Michael Crowley. 2019. "An Abrupt Move That Stunned Aides: Inside Trump's Aborted Attack on Iran." *New York Times*, Sept. 21, 2019. https://www.nytimes.com/2019/09/21/us/politics/trump-iran-decision.html.

Baraniuk, Chris. 2015. “Ashley Madison: ‘Suicides’ over Website Hack.” *BBC*, August 24, 2015. http://www.bbc.com/news/technology-34044506.

Barrett, Brian. 2019. “The Air Force Will Let Hackers Try to Hijack an Orbiting Satellite.” *Wired*, September 17, 2019. https://www.wired.com/story/air-force-defcon-satellite-hacking/.

Borghard, Erica D., and Shawn Lonergan. 2017. “The Logic of Coercion in Cyberspace.” *Security Studies* 26, no. 3: 452–481. https://doi.org/10.1080/09636412.2017.1306396

Chesney, Robert. 2019. “Crossing a Cyber Rubicon? Overreactions to the IDF’s Strike on the Hamas Cyber Facility,” *Lawfare*, May 6, 2019. https://www.lawfareblog.com/crossing-cyber-rubicon-overreactions-idfs-strike-hamas-cyber-facility.

Danzig, Richard. 2014. *Surviving on a Diet of Poisoned Fruit*. Washington, D.C.: Center for a New American Security. https://s3.amazonaws.com/files.cnas.org/documents/CNAS_PoisonedFruit_Danzig.pdf.

Futter, Andrew. 2018. *Hacking the Bomb: Cyber Threats and Nuclear Weapons*. Washington, D.C.: Georgetown University Press.

Gartzke, Erik and Jon R. Lindsay. 2017. “Thermonuclear Cyberwar.” *Journal of Cybersecurity* 3, no. 1 (March): 37–48. https://doi.org/10.1093/cybsec/tyw017.

Ghafur, S., S. Kristensen, K. Honeyford, G. Martin, A. Darzi, and P. Aylin. 2019. “A Retrospective Impact Analysis of the WannaCry Cyberattack on the NHS.” *Nature*, October 2, 2019. https://www.nature.com/articles/s41746-019-0161-6.

Goodin, Dan. 2015. “First Known Hacker-Caused Power Outage Signals Troubling Escalation.” *ArsTechnica*, January 4, 2015. arstechnica.com/security/2016/1/first-known-hacker-caused-power-outage-signals-troubling-escalation/.

Goodin, Dan. 2017. “Hackers Trigger Yet Another Power Outage in Ukraine.” *ArsTechnica*, January 11, 2017. https://arstechnica.com/security/2017/01/the-new-normal-yet-another-hacker-caused-power-outage-hits-ukraine/.

Gorman, Siobhan and Julian Barnes. 2011. “Cyber Combat: Act of War.” *Wall Street Journal*, May 31, 2011. https://www.wsj.com/articles/SB10001424052702304563104576355623135782718.

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

Greenberg, Andy. 2019. “New Clues Show How Russia’s Grid Hackers Aimed for Physical Destruction: A Fresh Look at the 2016 Blackout in Ukraine Suggests that the Cyberattack Behind it was Intended to Cause far More Damage.” *Wired*, September 12, 2019. https://www.wired.com/story/russia-ukraine-cyberattack-power-grid-blackout-destruction/.

Healey, Jason. 2019. “The Implications of Persistent (and Permanent) Engagement in Cyberspace.” *Journal of Cybersecurity* 5, no. 1: 1–15. https://doi.org/10.1093/cybsec/tyz008.

InfoSecurity. 2013. “Cyber-terrorism Shut Down Israel’s Carmel Tunnel.” October 28, 2013. http://www.infosecurity-magazine.com/news/cyber-terrorism-shut-down-israels-carmel-tunnel/.

Jensen, Benjamin, and Brandon Valeriano. 2019. *What Do We Know about Cyber Escalation? Observations from Simulations and Surveys*. Washington, D.C.: Atlantic Council. https://www.atlanticcouncil.org/wp-content/uploads/2019/11/What_do_we_know_about_cyber_escalation_.pdf.

Kahn, Herman. 1965. *On Escalation: Metaphors and Scenarios*. New York: Praeger.

Kostyuk, Nadiya, and Yuri M. Zhukov. 2017. "Invisible Digital Front: Can Cyber Attacks Shape Battlefield Events?" *Journal of Conflict Resolution* 63, no. 2: 317–347. https://doi.org/10.1177%2F0022002717737138.

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. https://doi.org/10.1093/cybsec/tyz007.

Leavitt, Lydia. 2011. "NASA Confirms Satellite Hacks in Congressional Advisory Panel." *Engadget*, November 2, 2011. https://www.engadget.com/2011/11/02/nasa-confirms-satellite-hacks-in-congressional-advisory-panel/.

Lin, Herbert. 2012. "Escalation Dynamics and Conflict Termination in Cyberspace." *Strategic Studies Quarterly* 6, no. 3 (Fall): 46–70. www.jstor.org/stable/26267261.

McElroy, Damien, and Ahmad Vahdat. 2013. "Iranian Cyber Warfare Commander Shot Dead in Suspected Assassination." *The Telegraph*, October 2, 2013. http://www.telegraph.co.uk/news/worldnews/middleeast/iran/10350285/Iranian-cyber-warfare-commander-shot-dead-in-suspected-assassination.html.

Morgan, Forrest E., Karl P. Mueller, Evans Medeiros, Kevin L. Pollpeter, and Roger Cliff. 2008. *Dangerous Thresholds: Managing Escalation in the 21st Century*. Santa Monica, CA: RAND.

Murgia, Madhumita. 2015. "Could Cyberattack on Turkey be a Russian Retaliation?" *The Telegraph*, December 18, 2015. http://www.telegraph.co.uk/technology/internet-security/12057478/Could-cyberattack-on-Turkey-be-a-Russian-retaliation.html.

Nakashima, Ellen. 2011. "U.S. Cyberweapons Had Been Considered to Disrupt Gaddafi's Air Defenses." *Washington Post*, October 17 2011. https://www.washingtonpost.com/world/national-security/us-cyber-weapons-had-been-considered-to-disrupt-gaddafis-air-defenses/2011/10/17/gIQAETpssL_story.html.

Newman, Lily Hay. 2018. "China Escalates Hacks against the US as Trade Tensions Rise." *Wired*, June 22, 2018. https://www.wired.com/story/china-hacks-against-united-states/.

Perlroth, Nicole. 2012. "In Cyberattack on Saudi Firm, U.S. Sees Iran Firing Back." *New York Times*, October 23, 2012. https://www.nytimes.com/2012/10/24/business/global/cyberattack-on-saudi-oil-firm-disquiets-us.html.

Reuters. 2012. "Suspected Cyber Attack Hits Iran Oil Industry." April 23, 2012. https://www.reuters.com/article/us-iran-oil-cyber/suspected-cyber-attack-hits-iran-oil-industry-idUSBRE83M0YX20120423.

Sanger, David and Mark Mazzetti. 2016. "U.S. Had Cyberattack Plan if Iran Nuclear Dispute Led to Conflict." *New York Times*, February 16, 2016. https://www.nytimes.com/2016/02/17/world/middleeast/us-had-cyberattack-planned-if-iran-nuclear-negotiations-failed.html.

Schneider, Jacquelyn. 2017. "Cyber and Crisis Escalation: Insights from Wargaming." https://pacs.einaudi.cornell.edu/sites/pacs/files/Schneider.Cyber%20and%20Crisis%20Escalation%20Insights%20from%20Wargaming%20Schneider%20for%20Cornell.10-12-17.pdf.

Tucker, Patrick. 2019. "The NSA Is Studying Satellite Hacking." *Defense One*, September 20, 2019. https://www.defenseone.com/technology/2019/09/nsa-studying-satellite-hacking/160009/.

Valeriano, Brandon, Benjamin Jensen, and Ryan Maness. 2008. *Cyber Strategy: The Evolving Character of Power and Coercion*. Oxford, U.K.: Oxford University Press.

United Nations. 2015. *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*. A/70/174. http://www.un.org/ga/search/view_doc.asp?symbol=A/70/174.

Zetter, Kim. 2015. "The NSA Acknowledges What We All Feared: Iran Learns from US Cyberattacks." *Wired*, February 10, 2015. https://www.wired.com/2015/02/nsa-acknowledges-feared-iran-learns-us-cyberattacks/.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Problems of Poison: New Paradigms and "Agreed" Competition in the Era of AI-Enabled Cyber Operations

Christopher Whyte
L. Douglas Wilder School of Government and Public Affairs
Virginia Commonwealth University

Abstract: Few developments seem as poised to alter the characteristics of security in the digital age as the advent of artificial intelligence (AI) technologies. For national defense establishments, the emergence of AI techniques is particularly worrisome, not least because prototype applications already exist. Cyber attacks augmented by AI portend the tailored manipulation of human vectors within the attack surface of important societal systems at great scale, as well as opportunities for calamity resulting from the secondment of technical skill from the hacker to the algorithm. Arguably most important, however, is the fact that AI-enabled cyber campaigns contain great potential for operational obfuscation and strategic misdirection. At the operational level, techniques for piggybacking onto routine activities and for adaptive evasion of security protocols add uncertainty, complicating the defensive mission particularly where adversarial learning tools are employed in offense. Strategically, AI-enabled cyber operations offer distinct attempts to persistently shape the spectrum of cyber contention may be able to pursue conflict outcomes beyond the expected scope of adversary operation. On the other, AI-augmented cyber defenses incorporated into national defense postures are likely to be vulnerable to "poisoning" attacks that predict, manipulate and subvert the functionality of defensive algorithms. This article takes on two primary tasks. First, it considers and categorizes the primary ways in which AI technologies are likely to augment offensive cyber operations, including the shape of cyber activities designed to target AI systems. Then, it frames a discussion of implications for deterrence in cyberspace by referring to the policy of persistent

engagement, agreed competition and forward defense promulgated in 2018 by the United States. Here, it is argued that the centrality of cyberspace to the deployment and operation of soon-to-be-ubiquitous AI systems implies new motivations for operation within the domain, complicating numerous assumptions that underlie current approaches. In particular, AI cyber operations pose unique measurement issues for the policy regime.

Keywords: *deterrence, persistent engagement, cyber, AI, machine learning*

1. INTRODUCTION

In recent decades, few technological developments have captured the attention and sparked the concern of national publics as much as those linked to artificial intelligence (AI). This might seem a remarkable and outlandish statement, given that, if prompted, the median consumer would likely be unable to identify that AI sits at the heart of everyday commercial services like Google's search engine or Amazon's marketplace. Nevertheless, the subject of AI has, since at least 2017, come to sit at the heart of prominent conversations about the future of human innovation and the changing shape of societal security.[1] Tech luminaries continue to expound the revolutionary potential of new machine learning and reasoning techniques which now easily solve those endemic issues of over-complexity that plague the conventional design and operation of digital systems. At the same time, leading voices – from Elon Musk to Max Tegmark and Steve Wozniak – increasingly refuse to disagree with doomsayers who claim that AI might, if mismanaged, lead to societal disaster.[2] Indeed, some are so concerned that they lean heavily into threat inflation, using extreme examples – such as the well-publicized threat of autonomous machine "slaughter bots" that, in a fictional future, catalyze societal breakdown as governments and private actors alike are empowered to kill opponents anonymously and at scale[3] – in an attempt to convince audiences of the stakes involved in getting AI "right".[4]

1 It should be noted that the topic of AI involved in the organization and application of military functions is not new, particularly in popular media. Instances of storytelling and more factual exploration can be found in film and written work stretching back through the early-mid 20th century.

2 See, among others, S. Hawking, S. Russell, M. Tegmark, and F. Wilczek, "Transcendence Looks at the Implications of Artificial Intelligence - But Are We Taking AI Seriously Enough?" *The Independent*, January 5, 2014; and Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Knopf, 2017).

3 Jessica Cussins. "AI Researchers Create Video to Call for Autonomous Weapons Ban at UN," *The Future of Life Institute*, accessed 28 November 2017, https://futureoflife.org/2017/11/14/ai-researchers-create-video-call-autonomous-weapons-ban-un/.

4 For an overview of expert opinion on AI, see Vincent C. Müller and Nick Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion," in *Fundamental Issues of Artificial Intelligence*, ed. Vincent C. Müller (Synthese Library; Berlin: Springer, 2016), 555-72.

Around the world, few entities are as focused on the impact that AI systems portend for security as national militaries. In the United States, political and military leaders have variously called for a "Third Offset" that leverages smart machine systems to outpace the capabilities of foreign adversaries in years to come.[5] Indeed, official strategy documents and the formal statements of such leaders today hold as a given fact what military practitioners and scholars generally take years to realize – that a new technology is changing the character of human warfare itself.[6] The resultant expectation, at least according to some, is that underlying AI processes will lead to an inevitable transformation in the bases of national power and will alter the constitution of security relationships between states in both strategic and operational terms.

This article contributes to the nascent literature on AI and national security activities by outlining the ways in which AI is likely to alter the shape of, and strategic calculations bound up in, interstate cyber conflict.[7] While there is a small-but-growing body of work on the potential of AI for affecting military and national power writ large, surprisingly few reports exist that attempt to problematize AI in the context of state

5 The "Third Offset" is a strategy intended to be used by the Department of Defense in the United States to counter and overcome advances being made by key peer competitors, such as China and Russia, in areas of military modernization and technology development. The term "Third Offset" refers to previous efforts to overcome perceived positional, military or technological advantages held by the Soviet Union during the Cold War, the first of which originated with the famed Project Solarium convened by President Dwight Eisenhower in the 1950s. Robert Work, "Remarks by Deputy Secretary Work on Third Offset Strategy," Brussels, Belgium, April 28, 2016, accessed 1 February 2018, https://www.defense.gov/News/Speeches/Speech-View/Article/753482/remarks-by-d%20eputy-secretary-work-on-third-offset-strategy/; Cheryl Pellerin, "Deputy Secretary: Third Offset Strategy Bolsters America's Military Deterrence," *DOD News* October 31, 2018, accessed 1 February 2018: https://www.defense.gov/News/Article/Article/991434/deputy-secretary-third-offset-strategy-bolsters-americas-military-deterrence/; and Katie Lange, "3rd Offset Strategy 101: What It Is, What the Tech Focuses Are," *DODLive* March 30, 2016, accessed 1 February 2018, http://www.dodlive.mil/2016/03/30/3rd-offset-strategy-101-what-it-is-what-the-tech-focuses-are/.

6 This point refers to the oft-cited manifestation of revolutions in military affairs (RMA) that dot human history. On the historical emergence of the RMA, see Dima Adamsky, *The Culture of Military Innovation: The Impact of Cultural Factors on the Revolution in Military Affairs in Russia, the US, and Israel* (Redwood City, CA: Stanford University Press, 2010) and Benjamin Jensen, "The Role of Ideas in Defense Planning: Revisiting the Revolution in Military Affairs," *Defence Studies*, forthcoming. On the distinction between a revolution in military affairs and military revolutions more broadly, see MacGregor Knox and Williamson Murray, eds., *The Dynamics of Military Revolution 1300-2050* (Cambridge: Cambridge University Press, 2001).

7 For a broad overview of the scope and dynamics of cyber conflict, see inter alia Brandon Valeriano and Ryan C. Maness, *Cyber War Versus Cyber Realities: Cyber Conflict in the International System* (Oxford University Press, USA, 2015); and Christopher Whyte and Brian Mazanec, *Understanding Cyber Warfare: Politics, Policy and Strategy* (Oxon and New York: Routledge, 2018).

competition online.[8] Moreover, what work does exist tends to involve only descriptive analysis of threat scenarios, pulling up short of considering how AI's augmentation of cyber capabilities – specifically the application of machine learning techniques to attack and defense – alters the dynamics of strategic engagement in the digital domain.[9] This article aims to act as a resource for those interested in thinking more clearly about how AI stands to alter the dynamics of both interstate conflict processes and cyber conflict processes more specifically.

In the sections below, I illustrate how AI-driven cyber attacks differ dramatically in their form from conventional digital threats. I then argue that, although such forms of attack are possible and likely beyond the digital domain, the centrality of cyberspace to the deployment and operation of soon-to-be-ubiquitous AI systems implies new motivations for operation within the domain. This dynamic, alongside the prospect of cyber offense and defense upgraded by AI, challenges several assumptions held by current strategies for cyber conflict prevention and should be a cause of significant concern for policymakers.

I proceed in three sections. First, I address the task of defining artificial intelligence as it is relevant to cyber operations. Here, I highlight the manner in which machine learning – technically a subfield of AI research that, according to many, now virtually demands consideration as its own technology – promises to affect many of the assumptions about operations in cyberspace that have been considered as standard among security practitioners and researchers for many years. I then describe the practical advancements to be expected with AI-driven cyber operations, as distinct from those that more substantially depend on the hacker in the loop, and categorize two particular forms of AI cyber attack. I then engage the topic of recent cyber conflict strategy and discuss AI developments in context, before concluding.

8 For the limited work to date on AI and strategic studies, see inter alia Benjamin M. Jensen, Christopher Whyte, and Scott Cuomo, "Algorithms at War: The Promise, Peril, and Limits of Artificial Intelligence," *International Studies Review* (2019); 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, 1-17 (IEEE, 2019); Kareem Ayoub and Kenneth Payne, "Strategy in the Age of Artificial Intelligence," *Journal of Strategic Studies* 39, no. 5-6 (2016): 793-819; Heather Roff, *Advancing Human Security Through Artificial Intelligence*, (Chatham House, May 2017, https://www.chathamhouse.org/publication/advancing-human-security-through-artificial-intelligence; Michael C. Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power," *Texas National Security Review* (2018); Kenneth Payne, *Strategy, Evolution, and War: From Apes to Artificial Intelligence*, (Georgetown University Press, 2018); Heather Roff, "COMPASS: A New AI-Driven Situational Awareness Tool for the Pentagon?" *Bulletin of the Atomic Scientists*, May 10, 2018, https:// thebulletin.org/compass-new-ai-driven-situational-awareness-tool-pentagon11816; Kenneth Payne, "Artificial Intelligence: A Revolution in Strategic Affairs?" *Survival 60*, no. 5 (2018): 7-32; Michael C. Horowitz, Gregory C. Allen, Elsa B. Kania, and Paul Scharre, "Strategic Competition in an Era of Artificial Intelligence," *Center for a New American Security* (2018); Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, et al. "The Malicious use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." *arXiv preprint arXiv:1802.07228* (2018).

9 See, for instance, Enn Tyugu, "Artificial Intelligence in Cyber Defense," in *2011 3rd International Conference on Cyber Conflict*, 1-11 (IEEE, 2011); or Mariarosaria Taddeo and Luciano Floridi, "Regulate Artificial Intelligence to Avert Cyber Arms Race," *Nature* 556 (2018): 296-298.

2. ARTIFICIAL INTELLIGENCE IN THE AGE OF CYBER CONFLICT

The label 'artificial intelligence' denotes a basket of technologies whose common attribute is the capability (or a set of capabilities) to simulate human cognition, particularly the ability of the human brain to adaptively reason, learn and autonomously undertake appropriate actions in response to a given environment.[10] In an even broader sense than is the case with all things "cyber," AI encompasses an immensely diverse landscape of technologies and areas of scientific development, from computer science to mathematics and neuroscience. The utilization of "AI" as a descriptor by many studies to describe new capabilities invariably risks, at least on some level, misleading readers by implying that artificial intelligence is best thought of as a relatively monolithic underlying technology whose design features will define future conflict. In reality, the implications of AI are best thought of in terms of unique interactions that will inevitably occur as an incredible array of potential smart machine systems is plugged into extant societal processes. This section attempts to contextualize the diverse form of what many simply generically refer to as "AI" and considers the implications for new techniques for the conduct of cyber conflict.

A. Machines That Reason, Learn and Act Autonomously

Machine cognition, which today substantially enables the function of most industrial sectors in advanced economies, has been a topic of significant interest to scientists and philosophers for the better part of two centuries. From Charles Babbage and Ada Lovelace to Alan Turing,[11] many of the greatest minds of the post-Industrial Revolution era have made their names by advancing societal thinking on the possibility of machines that can mimic how humans behave, move and think. More recently, the modern field of artificial intelligence – a term that emerged only in the early latter half of the 20th century among cybernetics and computer engineering researchers[12] – has its roots as a discipline in the substantial post-war work of minds like Marvin Minsky, Norbert Wiener and John von Neumann, who asked if, given the context of recent advances in computing, a machine might be made that could realistically simulate the higher functions of the human mind.[13] For such researchers, the challenge of machine intelligence lay in moving beyond the mere programmability of emerging computer constructs to building complex thinking systems capable of concept formation,

10 Jensen et al.(n 8) 10.

11 For contemporary description of such efforts, see inter alia Alan Turing, "Computing Machinery and Intelligence," *Mind* 49 (1950): 433-60; John von Neumann, *The Computer and the Brain*, (New Haven: Yale University Press, 1958); Nils J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements* (New York: Cambridge University Press, 2010); and Herbert Simon, "Artificial Intelligence: An Empirical Science," *Artificial Intelligence* 77, no. 2 (1995): 95-127.

12 Randolph Kline, "Cybernetics, Automata Studies, and the Dartmouth Conference on Artificial Intelligence," *IEEE Annals of the History of Computing* 33, no. 4 (October-December 2011): 5-16.

13 See Kline, ibid.; J. Moor, "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years," *AI Magazine* 27, no. 4 (2006): 87–91; and Bruce Buchanan, "A (Very) Brief History of AI," *AI Magazine* 26 (Winter 2005): 4.

environment recognition, abstract reasoning and self-improvement.[14] In the decades that have followed, of course, not only have such systems become commonplace in application to narrowly-defined societal functions, but competing schools of thought variously hold – for mathematical, neurological, evolutionary or computational reasons – that the future will see general learners whose ability to autonomously operate in the world matches and surpasses that of humans.

Today, AI, as applied broadly across areas of global society, is what researchers label "narrow" AI – not the "general" systems that are the focus of science fiction classics like *Terminator* or *I, Robot*, but limited applications of machine intelligence to discrete tasks.[15] Generally, though there is some crossover and some meaningful within-category differentiation, the technologies of AI might be thought of as existing across three main categories – (1) sensing and perception, (2) movement and (3) machine reasoning and learning.[16] Of these, the last is by far the one that is arguably most synonymous with AI as it is often portrayed in popular settings. In this category are a range of advances that encompass machines' abilities to interpret data, represent knowledge and understand information imbued with social meaning. By far the most significant area within this category is that of machine learning, the scientific study and development of approaches to pattern recognition and knowledge construction absent pre-programmed instructions on how to interpret data.[17] Machine learning is relatively simple to understand. Whereas conventional computing might involve the input of data to a non-learning algorithm in order to output some functional result, machine learning involves the input of both data *and* a desired result to an algorithm that infers, learns about a given issue represented in the data and then outputs another algorithm tailored to allow for intelligent engagement therewith.[18] In short, today's sophisticated AI techniques do not overwhelm computational challenges via the application of processing power so much as they more effectively study data to design a better process. In this way, AI promises to solve a traditional challenge in continuing to realize the promise of computers for human society – that the development of complex software to run on increasingly sophisticated systems means ever-growing demands on computer memory (both in storage and processing terms) and manifestations of human error in programming at scale. Machine learning compensates, not by building a better computer or catching those errors, but by allowing computers to sidestep such issues by programming and reprogramming themselves more efficiently.

14 McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. "A proposal for the Dartmouth Summer Research Project on Artificial Antelligence, August 31, 1955." *AI Magazine* 27, no. 4 (2006): 12-12; available at http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

15 Burton and Soare (n 8) 5.

16 Jensen et al.(n 8).

17 For an overview of machine learning, see Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning," *Nature* 521 (2015): 436-44. Also see V. Mnih et al., "Human-Level Control through Deep Reinforcement Learning," *Nature* 518 (2015): 529-33; and David Silver et al. "Mastering the Game of Go without Human Knowledge," *Nature* 550 (2017): 354-9.

18 For perhaps the most accessible description of machine learning at the point of operation, see Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, (Philadelphia, PA: Basic Books, 2015).

While machine learning involves these new processes and techniques for the direct mimicry of human cognition, the first two categories above include the technologies that are needed in order to allow machines to effectively move beyond internal processes to survey and operate within an environment. To some degree, of course, better sensing and perception are part and parcel of building better machine reasoning and learning algorithms. After all, effective mimicry of human cognition requires that such algorithms are able to interpret data and make inferences as a human might.[19] This involves an ability to consider language usage as a human might – i.e. more effective natural language processing (NLP)[20] – and a capability to construct and represent knowledge via ontological treatment. In this way, learner algorithms are able to move beyond simplistic statistical treatment of input data to identify concepts and connections that are sociological in nature.

Beyond the syntactic foundations of such advances in perception, however, much AI involves the development of new sensor systems that create data for algorithms to consume. Advances in camera systems and microwave sensors that allow for sophisticated text and imagery recognition via visual feeds, for instance, are critical to the function of new software that helps law enforcement more rapidly assess patterns in criminal behavior or traffic flow. At the same time, AI involves the construction of robotic systems that can more effectively gather data and can act as autonomous agents with the help of advanced learning software.[21] Though these areas of AI are less relevant for the discussion of cyber conflict in this paper, I address them further below.

B. Cyber Offense Enabled by AI

How might artificial intelligence augment or upgrade offensive cyber operations (OCO)? The conventional answer to such a question is simply that AI stands to make cyber attacks more powerful, to reduce the effectiveness of conventional defensive measures and to make powerful attacks more accessible for the median malicious online actor. More specifically, four prospective dynamics surrounding AI-enabled cyber offense seem worthy of note.

19 For a seminal description of perception as a component element of broader attempts to build deep learning and reasoning systems, see Nicola Jones, "The Learning Machines," *Nature* 505 (2014): 146-8.

20 For further information on NLP, see inter alia Stephen Deagelis, "The Growing Importance of Natural Language Processing," WIRED Magazine, February 2014, found at https://www.wired.com/insights/2014/02/growing-importance-natural-language-processing/; and Erik Cambria and Bebo White, "Jumping NLP Curves: A Review of Natural Language Processing Research," *IEEE Computational Intelligence Magazine* 9, no. 2 (May 2014): 48-57.

21 For further reading on intelligent machine vehicle systems, see *inter alia* Mario Gerla, Eun-Kyu Lee, Giovanni Pau, and Uichin Lee, "Internet of Vehicles: From Intelligent Grid to Autonomous Cars and Vehicular Clouds," *IEEE* (2014); and Alberto Broggi, Alex Zelinksy, Umit Ozguner, Christian Laugier, "Intelligent Vehicles," in *Springer Handbook of Robotics*, ed. B. Siciliano and O. Khatib, (Berlin, Heidelberg: Springer, 2016), 1627-56.

1) Attack Surface Analysis at Scale and Speed

First, AI programming portends a significantly increased threat to prospective cyber attack victims insofar as it enables analysis of the attack surface of targeted systems and victim entities at scale. This manifests at two levels. The first of these is the opportunity for malware to utilize incoming data obtained via infection of machines to probabilistically judge where and when further infection is likely to lead to some value return. An example of how such future AI-enabled malware might work can be found in the financial institution-targeting Trickbot malware encountered in just the past two years.[22] At the point of initial compromise, Trickbot functioned similarly to other worm-enabled malware seen since the mid-2010s. Once a foothold was established, however, multiple additional machines were compromised within minutes, without a clear pattern of target selection. Not only was the malware able to scale its attack at some speed; it also selected victims based on a "smart" analysis of prospective success in further infection. I place the word "smart" in quotation marks here because the malware was not truly utilizing the AI techniques baked into malware that many experts herald as coming in the near future, but rather was manually programmed to take more careful action. Nevertheless, the example stands as a case wherein rapid understanding of the attack surface of a target network led to an unusual strategy of infection – not every potential target was hit, only those with clear vulnerabilities in the form of outdated Server Message Block (SMB) services – that proved difficult and costly for defenders set up to handle less persistent threats.

The second manifestation of greater analysis of attack surfaces leading to increased digital insecurity lies in the wealth of data and metadata that either might be obtained via traditional intelligence methods or are already available from criminal sources. The more data available to malicious actors interested in leveraging the advantages of AI for cyber aggression, the more capable the techniques employed might be. The future may very well hold cyber campaigns of either a criminal or a political nature which would be substantially informed by the wealth of data that might be made available to attackers for analysis. The gold standard of AI-enabled OCO, particularly those that target broad populations or large institutions, is one substantially designed by learning systems that infer lateral approaches to targets – and, in some cases, rapidly and autonomously undertake malicious action informed by such inference – with relatively low risk of detection or mitigation. Indeed, this threat of attack surfaces under sophisticated machine intelligence scrutiny is one of the core challenges that promises to impact current thinking on cyber conflict strategy and signaling. I return to this point in detail below.

22 For a description of the episode in context, see Cyber-Attacks, AI-Driven. "The Next Paradigm Shift." Also see Lior Keshet, "An Aggressive Launch: TrickBot Trojan Rises with Redirection Attacks in the UK," *Security Intelligence* (2016); and Darrel Rendell, "Understanding the Evolution of Malware," *Computer Fraud & Security* 2019, no. 1 (2019): 17-19.

2) Technique Adaptation

A second dynamic surrounding AI-enabled cyber offense is the ability of malware to autonomously select from a toolkit of options for further spread. Malware that is inserted into a machine might undertake environmental analyses and determine that another technique is more suited to attaching new victims than was the particular exploit involved in the initial compromise. Here, the shape of the AI-enabled cyber attack is not very different from the sophisticated software often employed by state security institutions or other advanced persistent threat actors (APTs). It is simply a more accessible, automatable method for empowering hackers of all stripes to utilize tools smart enough to fit variable elements of an attack toolkit into a diverse attack surface.

3) Adversarial Tactical Adaptation

Third, the threat of cyber offense upgraded by AI is also a type of malware that is able to adjust its own strategy of approach as operations are underway. Different from having a simple ability to assess potential targets and select appropriate methods of approach, AI programming will allow malware to alter its tactics in line with mission parameters as it learns more and more about the environment in which it is operating – and the defenders and users that populate that environment. Faced with diverse defense efforts across a diverse multi-network attack surface, a sophisticated AI-enabled attack on defense infrastructure could, for instance, determine that the rapid promulgation most advisable for one institution – say, a research laboratory – would be associated with greater risks of detection if executed against another target – say, a military base of operations. In such circumstances, the same piece of malware might be able to select an alternative approach, such as hiding or going "slow-and-low" in its effort to compromise machines and exfiltrate information. In this way, AI-enabled malware presents as an adversarial threat that functions even when – indeed, arguably especially when – robust defender efforts are apparent.

4) Multiple Mindsets

Finally, experts are concerned that AI-enabled malware will be able to analyze victim networks at scale and act autonomously to attack in ways that maximize opportunities for further compromise. A sub-element of the ability of AI-enabled malware to change tactical approach even beyond the point of victim identification and promulgation is the opportunity for multi-purpose malware that might change its own task or learn new tasks within the context of an existing operation. AI programming will allow sophisticated malware to learn about the defensive environment and compartmentalize lessons learned, such that alternative "mindsets" can drive activity where mission parameters are deemed to have changed (for example, upon discovery of a supervisory control system or where information has been retrieved and the task becomes one of exfiltration).

C. Cyber Artificial Intelligence Attacks: Threat Types

Naturally, if the potential underlying artificial intelligence for cyber offense can be summed up as greater adaptability, rapidity and opportunity for unexpected malicious behavior, then something similar can be said of the potential of AI-enabled cyber defenses. And indeed, it would be unfair to broach any discussion of the prospective impact of AI on cyber conflict without considering that the new learning, reasoning and sensing techniques will also come to – and already have begun to – undergird the efforts of defenders. Just as AI stands to augment and enhance the offense, so too will it become a necessity for those humans in the loop whose conventional perimeter, simulative and dissimulative defenses become the fodder from which adversarial attack AI builds better offensive routines.[23] Even here, however, it would be disingenuous to suggest that the AI arms race in cyber capabilities can be boiled down to tit-for-tat improvements in the relative capacities of those on the offense or defense. There are complex challenges facing those on the defense in the form of cyber artificial intelligence attacks (CAIA), which are attacks that seek to take advantage of approaches to system operations and defender routines in practice in order to subvert the legitimate functionality thereof.[24] In other words, CAIA essentially constitutes attacks against the AI itself that will increasingly come to underwrite cyber conflict processes. Such attacks might fall into two categories.

1) Input Attacks

Input attacks are those forms of contestation that seek to fundamentally mislead an AI system and skew the efforts of that system to classify patterns of activity.[25] If the expectations of a model designed by a learning AI program can be subverted, new space opens for unique, hard-to-predict exploits. Notably, input attacks do not involve attacking the code of AI systems or plugins itself; rather, the point of input attacks is deception that aims to control – or, at least, partially shape – how an AI system is "thinking" about a given issue or functional challenge. In this way, input attacks are best thought of as counter-command and control (counter-C2) warfare.[26]

Input attacks are highly varied in their form and can functionally be a great many things. This is because input attacks are defined by the function and deployment of the models they target. They might even involve physical activities in aid of cyber outcomes. For instance, a hypothetical re-running of the Stuxnet attack on Iran's

23 For discussion of simulation as an element of strategic interactions in cyberspace, see Erik Gartzke, and Jon R. Lindsay, "Weaving Tangled Webs: Offense, Defense, and Deception in Cyberspace," *Security Studies* 24, no. 2 (2015): 316-48.

24 The term "cyber artificial intelligence attacks" is inspired by its recent usage in Marcus Comiter, *Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It*, Belfer Center for Science and International Affairs, Harvard Kennedy School, August 2019.

25 Ibid. 19.

26 See Norman B. Hutcherson, *Command and Control Warfare. Putting Another Tool in the War-Fighter's Data Base*. No. AU-ARI-94-1, (Air Univ Maxwell AFB AL Airpower Research Inst, 1994); and Jeffrey A. Harley, *The Role of Information Warfare: Truth and Myths*, (Naval War Coll Newport RI Joint Military Operations Dept, 1996).

uranium enrichment facility at Natanz in which the defenders employed AI in the defense of internal networks might necessitate a nascent phase wherein the malware lay dormant vis-à-vis its core purpose and undertook secondary actions to install internal methods of subverting key defender system functions. At the same time, however, the malware might also benefit from input attacks undertaken by human intelligence assets. For instance, a piece of tape placed on one or more computer monitors on-site could conceivably trick security cameras into believing that those monitors were always on. Those cameras would not then flag an anomaly when malware turned a machine on during a period of inactivity.

2) Poisoning Attacks

In contrast with input attacks, poisoning attacks are activities that fundamentally seek to compromise the AI programming employed in enemy systems.[27] In the Stuxnet *redux* example above, such an attack on the part of the malware involved might, among other things, involve gradually increasing traffic volume to certain machines during non-peak hours. Therein lies the primary way in which AI systems are "poisoned" – the manipulation of data that such systems are trained upon so that the model learned by the target system does not accurately reflect reality. In poisoning an AI system, attackers in essence create backdoors via which further offensive action might be taken. This can, naturally, take a number of formats. An attacker might "train" a defending model to be oblivious to specific forms of anomalous behavior. Likewise, a system might be persuaded to fail or trigger some otherwise unrelated – but useful – process at a particular time when a certain action, such as a diagnostic scan, is taken.

3) Thinking About CAIA at Scale

It is tempting to primarily think of AI-enabled attacks as targeting the functionality of AI systems which defenders increasingly rely on to undertake security actions. However, the implications of CAIA for national security apparatuses go beyond such considerations. Specifically, the problem of poison for modern security institutions exists in such a way that the cyber-specific context implied in the threat type descriptions above constitutes only one element of the challenge. Given the coming proliferation of AI across military functions, security planners face the threat of skewing from nigh-uncountable sources. If adversary militaries wish to skew North Atlantic Treaty Organization (NATO) analytics, they might utilize conventional military deception methods – such as deploying decoy vehicles during military maneuvers to mislead NATO forces about the normal scale and dispersion of adversary forces – to do so as easily as they might tamper with training data via cyber means. Thus, it would be at least partially disingenuous to argue here that the augmentation of cyber conflict processes by AI constitutes a unique-to-the-domain coming transformation.

27 See Comiter (n 24) 28.

3. SHAPING BEHAVIOR IN AN AGE OF ADVERSARIAL LEARNING[28]

What *is* particularly unique about the intersection of artificial intelligence and cyber conflict processes, however, is that the centrality of cyberspace to the deployment and operation of soon-to-be-ubiquitous AI systems implies new motivations for operation within the domain. The prospect of subverting AI-driven security functions – in particular, the prospect of fundamentally poisoning the deliberative and operational bases of important national security establishment functions – provides incentive for operation in cyberspace beyond in-domain effects and outcomes. On the one hand, cybersecurity experts might expect an intensification of cyber conflict and criminal activities around the world based on near-term adoption of advancing AI programming that promises rapid adaptability and sophistication without either major investment or the need for major human presence in the loop. On the other hand, the same experts might expect an intensification of such activities because CAIA will so clearly often involve effects beyond the domain (e.g. cyber operations that are not operationally focused on some digital compromise so much as they are intended to affect real-world approaches to risk management, strategic assessment and resultant military deployments, financial outlays, etc.).

In the remaining section of this paper, I consider the implications of AI-augmented cyber attacks and CAIA for current strategic approaches to the mitigation of cyber conflict. Specifically, I describe the strategy of forward defense based around the dynamics of persistent engagement between adversaries in the domain that now constitutes American Title 10 approaches to operation online and suggest several core problems that either intensify or newly manifest in an era of large-scale proliferation of AI in cyber. The focus on U.S. strategy is intentional, as changes to America's force posture in the fifth domain represent the concrete edge of efforts to adapt prevailing approaches to cyber conflict in the context of both intensifying digital interference since 2010 and the failing applicability of legacy security concepts to the challenge. The dynamics of AI-augmented cyber conflict and the related questions that must be addressed vary beyond the scope of such singular focus, of course. But national contextualization allows for more in-depth exploration and produces analytic outcomes generalizable beyond the specific case.

A. Persistent Engagement and Defending Forward

In 2018, as it was elevated to the status of unified combatant command within the U.S. military, Cyber Command promulgated a new strategic vision centered around the

28 The phrase "adversarial learning" is a common one utilized by computer scientists to describe how machine learning algorithms are capable of adapting to hostile operational environments by crystalizing alternative – rather than combative – approaches to operation. See *inter alia* Daniel Lowd and Christopher Meek, "Adversarial Learning," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, ACM (2005): 641-47; and Pavel Laskov and Richard Lippmann, "Machine Learning in Adversarial Environments," *Machine Learning* 81 (2010): 115-19.

concept of persistent engagement.[29] To put the concept and the strategy that emerges therefrom bluntly, persistent engagement means that Cyber Command intends to be everywhere, constantly maintaining presence and employing necessary tools against America's adversaries in networks wherever they might be found. The strategy pushes back against strategy as practiced in the past by both American administrations and allies, wherein operations were based on the political desire to mitigate cyber risk principally via norm development and through deterrent efforts that stemmed substantially from Cold War postures.[30]

In terms of the strategic logic of engagement in the domain, the persistent engagement strategy largely emerges from the work of Harknett and Fischerkeller in their time as scholars attached to Cyber Command. The authors argued that the unique character of cyberspace means that traditional deterrent approaches are doomed to failure.[31] Given that deterrence involves strong demonstrations of defense or meaningful statements of punishment following attacks, the prospects for developing a sustainable deterrent posture online are limited.[32] It is extremely difficult to demonstrate defensive capabilities at the scale demanded by a national cyber deterrent strategy, and punishment rarely works in the way it is intended. Communicating specific meaning in retaliation is difficult, particularly where the diversity of activities that constitute cyber conflict is immensely high. Moreover, response options are often not ready to go in the timeframe required by policymakers who seek to deter. Further, conceptual agreement on the significance or role of certain elements of the domain is not easy to come by, with poor understanding of what might be meant – if anything – by sovereignty online being a hallmark of the digital world.

The result is an alternative strategy – persistent engagement – that emphasizes "defending forward." This posture involves cyber forces operating beyond government and domestic networks to actively contest enemy activities aimed at harming national security or other national interests. Such operations, it is argued, can avoid escalation by embracing the doctrine of selective engagement and can be designed specifically to scale tactical efforts into strategic gains. In doing so, the idea is that the behavior of adversaries can be shaped and the scope of what is deemed to be appropriate

29 Department of Defense, *National Cyber Strategy of the United States of America*, 2018.

30 Nakasone, Paul M. "An Interview with Paul M. Nakasone," *Joint Forces Quarterly* (2019). https://ndupress.ndu.edu/Portals/68/Documents/jfq/jfq-92/jfq-92_4-9_Nakasone-Interview.pdf.

31 Fischerkeller, Michael P., and Richard J. Harknett. "Deterrence is not a credible strategy for cyberspace." *Orbis* 61, no. 3 (2017): 381-393.

32 For the broad literature on deterrence in cyberspace, see *inter alia* Libicki, Martin C. *Cyberdeterrence and cyberwar*. (Rand Corporation, 2009); Lupovici, Amir. "Cyber warfare and deterrence: trends and challenges in research." *Military and Strategic Affairs* 3, no. 3 (2011): 49-62; Crosston, Matthew D. "World gone cyber MAD: How 'mutually assured debilitation' is the best hope for cyber deterrence," *Strategic Studies Quarterly* 5, no. 1 (2011): 100-116; Jensen, Eric Talbot. "Cyber deterrence." *Emory Int'l L. Rev*. 26 (2012): 773; Denning, Dorothy E. "Rethinking the cyber domain and deterrence" (2015); Iasiello, Emilio. "Is cyber deterrence an illusory course of action?" *Journal of Strategic Security* 7, no. 1 (2014): 54-67; and Tor, Uri. "'Cumulative Deterrence' as a New Paradigm for Cyber Deterrence," *Journal of Strategic Studies* 40, no. 1-2 (2017): 92-117.

competition can be made known.[33] The resultant condition should, it is hoped, be one of "agreed competition" wherein the bounds of cyber conflict that are deemed to be acceptable can be consistently made known and where the worst excesses of digital insecurity for states might be avoided by the institution of precise conditions of case-by-case deterrence.[34]

B. Basic Challenges of AI for Persistent Engagement

Thinking effectively about the problem of poison for cyber conflict processes – particularly as a subset of all national security processes – is tricky, in that we have to fundamentally think about learning as it manifests in two different settings: in the organizational setting and in the construction of AI systems. It is not simply enough to consider the impact of rapid learning techniques for cyber conflict as we understand it today, though that approach to thinking about the problem of AI in this area *does* suggest some obvious challenges to be faced by prevailing strategy.

Above almost all other implications, broad-scoped upgrading of "conventional" cyber techniques portends a narrowing of the space within which adversaries might undertake cost-benefit calculations and come to believe that the benefits of further action are outweighed by the costs that might be imposed in the domain by forward defenders. Simply put, if smart tools exist that can more reliably avoid detection, take lateral routes to targets, or scale effects much more quickly than is the norm today, then adversaries are likely to exhibit increased willingness to continue operations under circumstances where they would not previously have done so. Especially given that the stakes of defection from agreed conditions of competition are not typically very high in political terms, this contraction of that space, wherein persuasion is argued to be possible under a doctrine of persistent engagement, ostensibly makes meaningful signaling even trickier from situation-to-situation. Likewise, at the most basic level, the proliferation of relatively robust abilities to achieve effects in the digital domain via lateral action – i.e. action that takes indirect, harder-to-predict pathways toward targets and outcomes – suggests that we might see recurrent incidents in areas where a threat had previously been thought to have been realized and countered in some form.[35]

33 Fischerkeller, Michael P., and Richard J. Harknett. "Persistent Engagement, Agreed Competition, Cyberspace Interaction Dynamics and Escalation." Orbis (Summer 2017) 61, no. 3 (2018): 381-393.

34 See *inter alia* Defense Science Board, Department of Defense. 2017. "Task Force on Cyber Deterrence." Defense Science Board, 3, 4. https://apps.dtic.mil/docs/citations/AD1028516; Bolton, John. 2018. "Transcript: White House Press Briefing on National Cyber Strategy - Sept. 20, 2018." Washington DC (September 8). Available at https://news.grabien.com/making-transcript-white-house-press-briefing-national-cyber-strategy.

35 This point references the oft-cited framing of cyber conflict history in the West as emerging via a series of realization episodes that have prompted a series of institutional and doctrinal adaptations over the past three decades. See Jason Healey (ed.), *A Fierce Domain: Conflict in Cyberspace, 1986 to 2012*, Cyber Conflict Studies Association, 2013.

It is perhaps most particularly worth noting that AI-enabled cyber conflict adds a new dimension to the traditional perception problem experienced in cyberspace, where attribution of intent or agency is particularly difficult at the point of threat detection and analysis.[36] Where a probing attack or some other action is detected, it is rare that the investigator is able to discern between run-of-the-mill adversary efforts to conduct espionage or some attacking action. In the near term, another possibility is that cyber actions may be not linked with either espionage or direct attack, but rather with attempts to interfere with the function of AI programming.[37] The particular danger here is that such attempts may involve activities that are even less clearly discernible as aggressive or not than is the case with espionage activities.

C. The Learning Problem

Beyond the basic challenges to the strategy of persistent engagement posed by the intensification of cyber conflict driven by the adaptability and rapidity brought by AI, of course, policymakers and practitioners must inevitably grapple with increasing uncertainty around the state of common knowledge between actors in the domain. The perception dynamic described above, for instance, is uniquely concerning for current strategic thinking on cyber conflict management, insofar as cyberspace is likely to be the domain of political activity most central to efforts to poison or otherwise interfere with AI systems. In a future where conflict involves broad-scoped efforts to manipulate the construction and operation of AI systems attached to myriad societal functions, cyberspace constitutes the primary highway via which such shaping efforts will likely flow. Moreover, state interest in operations of a poisoning nature via cyberspace is likely to grow over time as opportunities proliferate for the manipulation of processes that underlie strategy development, force posture determination and more.[38] Both of these points mean that strategic efforts to constrain adversaries' cyber actions relative to in-domain considerations may fail simply because they are not effectively armed with appropriate assumptions about the motivations of actors to operate online.

More broadly, the advent of narrow AI baked into most functional elements of a state's national security apparatus implies an enduring tension in the conduct of persistent operations intended to shape adversary behavior. All else being equal, the existence of robust AI systems on the part of foreign adversaries implies a learning problem –

36 See *inter alia* Nicholas Tsagourias, "Cyber Attacks, Self-Defence and the Problem of Attribution," *Journal of Conflict and Security Law* 17, no. 2 (2012): 229-44; Jon R. Lindsay, "Tipping the Scales: The Attribution Problem and the Feasibility of Deterrence Against Cyberattack," *Journal of Cybersecurity* 1, no. 1 (2015): 53-67; and Thomas Rid and Ben Buchanan, "Attributing Cyber Attacks," *Journal of Strategic Studies* 38, no. 1-2 (2015): 4-37.

37 This issue lies at the heart of what Buchanan labels the "cybersecurity dilemma." See Ben Buchanan, *The Cybersecurity Dilemma: Hacking, Trust, and Fear Between Nations* (Oxford: Oxford University Press, 2016).

38 This assertion is quite arguably backed by work that demonstrates in both quantitative and qualitative terms an increasing turn towards political warfare as an adjunct of cyber conflict, in line with the proliferation of digital services and social platforms that undergird major societal functions. See, for instance, Brandon Valeriano, Benjamin M. Jensen, and Ryan C. Maness, *Cyber Strategy: The Evolving Character of Power and Coercion* (Oxford: Oxford University Press, 2018).

the more security institutions operate to shape behavior, the more those adversaries *should* be empowered to understand and overcome such strategies. After all, much as in the case of Generative Adversarial Networks (GANs) that study the actions of AI models in order to continually improve offensive capabilities,[39] AI-enabled cyber forces presented with unique patterns of behavior-shaping attack from abroad will naturally undergo a process of adversarial learning where foreign action does not bound the shape of acceptable behavior so much as define the criteria under which future aggression is probabilistically less likely to induce some cost. Particularly given the incentive described above towards the use of AI-enabled software agents that have dramatically higher track records of success – given their adaptability – than non-AI-enabled versions, the commonplace existence of such systems seems likely to work against the development of static norms of behavior.

Finally, the result of an emergent era in which AI-driven adversarial learning is the key feature of interstate interactions online is a perpetual challenge of validation. In recent scholarship, there have already been some discussions about the challenges involved in applying relevant metrics to the strategy of persistent engagement such that defense practitioners might determine its effectiveness.[40] Such challenges multiply, given the AI-ification of cyber conflict processes and the problem of poison as a regular feature of operation in the domain. Whereas analysis of broad patterns of activity might otherwise offer some indication as to the effectiveness of forward defensive efforts aimed at dissuading particular adversary behaviors, such metrics may not apply in a significant fashion in an era where counter-action from foreign peers is not expected to be tit-for-tat, but rather entirely alternative in approach. In other words, where the paradigm of operation shifts from in-kind engagement – even if that engagement emerges from an admittedly diverse toolkit – to an imperative of lateral approach and misdirection, attempts to validate current strategic processes seem likely to be ineffective beyond simplistic analysis of major event incidence.

4. IMPLICATIONS FOR STRATEGIC THINKING

The purpose of this article is to contribute to the nascent literature on AI and national security activities by outlining the ways in which AI is likely to alter the shape and strategic calculations bound up in interstate cyber conflict. It is hoped that the sections above can act as a resource for those interested in thinking more clearly about how AI stands to alter the dynamics of both interstate conflict processes and cyber conflict processes. Naturally, a substantial part of the effort made herein has been definitional. Indeed, it is from the categorization of different threat forms linked to the

39 Vincent, James. "Deepfake Detection Algorithms Will Never Be Enough." *The Verge* (2019).

40 See, for instance, Jason Healey and Neil Jenkins, "Rough-and-Ready: A Policy Framework to Determine if Cyber Deterrence is Working or Failing," in *2019 11th International Conference on Cyber Conflict (CyCon)*, vol. 900 (IEEE, 2019): 1-20.

augmentation of cyber conflict processes by AI models and systems that the primary argument of this paper emerges – that the centrality of cyberspace to the deployment and operation of soon-to-be-widespread AI systems implies new motivations for operation within the domain. The implications thereof for current cyber conflict strategies – particularly those being worked on by Western defense establishments – are numerous and remain to be assessed in full as literature on the subject is developed in the future. Nevertheless, some immediate takeaways are apparent.

First, strategic planners and policymakers must recognize from the start that there are two levels of challenge when it comes to AI augmentation of cyber conflict processes. At the first level, AI promises to reduce the window in which it may be possible to shape competition in cyberspace in favorable terms. At the second, AI intensifies and adds a new dimension to the challenges of validity and attribution already present in cyber operations. Simply put, given the opportunities for poisoning by soon-to-be-ubiquitous AI models at work in security apparatuses, how can defenders really know what they think it is they know about the integrity of their systems? At the strategic level, given that broad-scoped attempts to shape competition between AI-enabled adversaries are likely to empower opponents via a process of adversarial learning, how can policymakers and military practitioners really know what they think it is they know about strategic conditions?

Second, because of the various challenges bound up in effectively deploying AI for national security purposes, the effectiveness thereof is likely to be bound up in the approach organizations take to trusting their AI systems and to managing the interaction of human and machine operators.[41] Much of what has been discussed in the sections above involves – to at least some degree – the problem of ghosts in the machine, where it is human assumptions present in the code of machine intelligence systems that form the true problem for effective deployment for national security purposes. While such problems are arguably unavoidable as we move toward more common employment of AI than is the case today, it seems likely that protocols for keeping humans in the loop at critical junctures are part of the solution to problems of (either malicious or self-inflicted) poison.

Finally – and perhaps most significantly – it seems clear that, in the forthcoming era of AI-enabled contestation in world affairs, strategy development, assessment and validation must emerge significantly from cross-domain understanding of the strategic motivations of adversaries. If cyberspace is not only a domain wherein unique forms of contestation and signaling can take place, but is also the most significant terrain over which actions can be taken to affect processes that underlie all areas

41 This is not a thus-far uncommon argument made by scholars of cyber conflict. See, for instance, Rebecca Slayton. "What is the Cyber Offense-Defense Balance? Conceptions, Causes, and Assessment," *International Security* 41, no. 3 (2017): 72-109.

of modern society, then strategic planners would do well to build from assumptions that move beyond simple logic-of-the-domain characterizations of digital affairs. As has previously been argued in both implicit and explicit terms,[42] cyber conflict so often manifests in aid of non-digital contestation that we would do well to couch our analyses in terms of the logic of conflict processes *other* than cyber. This stands to be especially the case with artificial intelligence, not least given the fact that the targeting of AI for security purposes is so likely to be significantly tied to use of the computer and Internet systems upon which such programming must inevitably run.

42 See, for instance, Christopher Whyte, "Dissecting the Digital World: A Review of the Construction and Constitution of Cyber Conflict Research," *International Studies Review* 20, no. 3 (2018): 520-32; and Jon R. Lindsay, "Stuxnet and the Limits of Cyber Warfare," *Security Studies* 22, no. 3 (2013): 365-404.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

The Next Generation of Cyber-Enabled Information Warfare

Kim Hartmann
Conflict Studies Research Centre
kim.hartmann@conflictstudies.org.uk

Keir Giles
Conflict Studies Research Centre
keir.giles@conflictstudies.org.uk

Abstract: Malign influence campaigns leveraging cyber capabilities have caused significant political disruption in the United States and elsewhere; but the next generation of campaigns could be considerably more damaging as a result of the widespread use of machine learning.

Current methods for successfully waging these campaigns depend on labour-intensive human interaction with targets. The introduction of machine learning, and potentially artificial intelligence (AI), will vastly enhance capabilities for automating the reaching of mass audiences with tailored and plausible content. Consequently, they will render malicious actors even more powerful.

Tools for making use of machine learning in information operations are developing at an extraordinarily rapid pace, and are becoming rapidly more available and affordable for a much wider variety of users. Until early 2018 it was assumed that the utilisation of AI methods by cyber criminals was not to be expected soon, because those methods rely on vast datasets, correspondingly vast computational power, or both, and demanded highly specialised skills and knowledge. However, in 2019 these assumptions proved invalid, as datasets and computing power were democratised and freely available tools obviated the need for special skills. It is reasonable to assume that this process will continue, transforming the landscape of deception, disinformation and influence online.

This article assesses the state of AI-enhanced cyber and information operations in late 2019 and investigates whether this may represent the beginnings of substantial and dangerous trends over the next decade. Areas to be considered include: social media

campaigns using deepfakes; deepfake-enabled CEO fraud; machine-generated political astroturfing; and computers responding to the emotional state of those interacting with them, enabling automated, artificial humanoid disinformation campaigns.

Keywords: *deepfake, disinformation, information warfare, malign influence, artificial intelligence, machine learning, emotional modelling*

1. INTRODUCTION

The year 2019 saw rapid developments in the use of machine-learning techniques to assist and amplify malign influence campaigns. Early in the year, "Katie Jones" was the first publicly identified instance of a deepfake face image used in a social media campaign.[1] By December, this technique had gone mainstream, with mass use in a campaign to influence US politics.[2] It is highly likely that this trend will continue in information operations, and as a result may transform the techniques, capabilities, reach and impact of information warfare.

Advances in artificial intelligence (AI) and the increasing availability of manipulation software usable by laymen have made the creation of convincing fake audio and video material relatively easy. The rapid spread of such material through social media and a lack of sufficient validation methods in cyberspace have resulted in the emergence of a potentially very powerful weapon for information operations. The speed of progress in this field is such that while deepfakes were not relevant for the 2016 US presidential election – at present the most prominent case study of cyber-enabled hostile interference in an election campaign – in 2020 they are widely regarded as a significant danger.

Until this point, malign influence and disinformation campaigns have primarily been operated and directed manually, or with the assistance of relatively crude and simple bots that are not able to interact convincingly with human targets or generate strategic long-term engagement. The design, production and dissemination of false material have been performed by human operators. But the trend of utilising AI methods to compose manipulated or fake material observed during 2019 indicates that it is possible to automate the processes needed to successfully operate disinformation

1 Keir Giles, Kim Hartmann, and Munira Mustaffa, *The Role of Deepfakes in Malign Influence Campaigns*, (Riga: NATO STRATCOM COE, 2019), https://www.stratcomcoe.org/role-deepfakes-malign-influence-campaigns.

2 Davey Alba, "Facebook Discovers Fakes that Show Evolution of Disinformation", *The New York Times*, 20 December 2019, https://www.nytimes.com/2019/12/20/business/facebook-ai-generated-profiles.html.

campaigns. In particular, this is because the level of sophistication of AI reached in a data processing and reasoning application context is different to AI in other fields. This type of AI may be considered as in between what are referred to as "strong" and "weak" AI. "Weak" AI is already available for generating specific output material or discrete tasks involved in disinformation operations, when the prerequisites and other inputs required to automate and generalise these tasks are already given. Currently these AI applications remain field-specific and hence cannot be considered as "strong" or true AI; however, with the appropriate supply of prerequisites and input data required to automate and generalise these tasks, their capabilities are much higher than the average "weak" AI already observed today.

While AI is still immature in many application scenarios, the technology has made significant steps in the specific areas of data analysis, classification, creation and manipulation, with a significant rise in the achievable output due to the availability of high-quality data and data processing routines (big data) as well as CPU power and memory capacities. While it is still difficult for AI systems to adapt to the real world, cyberspace – being an artificially generated domain constructed around pure data and communication – is their natural environment.

Most societies are still relatively accustomed to viewing audio and video recordings as indisputable evidence of reality. Images, video and audio recordings have played a major role in documenting our recent history and our trust in these recordings has shaped our perception of reality. Without modern media and our trust in them, our history is likely to have been different. An example is the release of the former US President Richard Nixon's "smoking gun" tape, which eventually led to a change of power in the United States. Had this tape not existed, or had it not been trusted, history could have taken a completely different course.

In facing an era of artificially generated images, audio and video recordings, we are also confronted with the risk of real events being falsely claimed to be fake. As we currently do not have sufficient technologies to guarantee the authenticity of material being displayed, proving a falsely-claimed fake to be real may be even more challenging than the reverse. The effect of such false claims, especially in a political context, may be immense.

We have entered an era in which we depend heavily on audio and video materials as information resources while at the same time being confronted with the fact that this material can no longer be fully trusted. Although there have always been individuals who consider historic events such as the Holocaust, the moon landings or even the 9/11 terror attacks to be fictions despite multiple media evidence, current studies indicate that the number of individuals distrusting facts is rising rapidly due

to the emergence of deepfake technology.[3] The erosion of trust in objective truth is accelerated by the ease with which apparently reliable representations of that truth can be fabricated; and augmented by the secondary effect of reduced trust in mainstream media, which neutralises their role in providing facts, informing the public and thus stabilising democratic processes. This plays directly into the hands of organised disinformation campaigns. A 2019 JRC Technical Report on the Case Study of the 2018 Italian General Election, published by the European Commission, indicated a correlation between distrust in media and a higher susceptibility to disinformation.[4]

Members of the US Congress have requested a formal report from the Director of National Intelligence on deepfakes and the threats they pose.[5] US senators Marco Rubio, member of the Senate Select Committee on Intelligence, and Mark Warner, Chairman of the Senate Select Committee on Intelligence, have urged social media companies to develop standards to tackle deepfakes, in light of foreign threats to the upcoming US elections. They note that: "If the public can no longer trust recorded events or images, it will have a corrosive impact on our democracy".[6]

Meanwhile, by 2018 the US defence research agency DARPA had spent 68 million US dollars on a four-year programme developing digital forensics to identify deepfakes. However, there is a concern that the defending side in combating deepfakes will always be at a disadvantage. According to Hany Farid, a digital forensics expert at Dartmouth College: "The adversary will always win, you will always be able to create a compelling fake image, or video, but the ability to do that if we are successful on the forensics side is going to take more time, more effort, more skill and more risk".[7]

3 Karen Hao, "The biggest threat of deepfakes isn't the deepfakes themselves - The mere idea of AI-synthesized media is already making people stop believing that real things are real", *MIT Technology Review*, 10 October 2019, https://www.technologyreview.com/s/614526/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/; Simon Kuper, "The age of scepticism: from distrust to 'deepfake'", *Financial Times Magazine*, 18 October 2018. https://www.ft.com/content/2fc9c1fa-d1a2-11e8-a9f2-7574db66bcd5.

4 Massimo Flore, Alexandra Balahur, Aldo Podavini, Marco Verile, "Understanding Citizens' Vulnerabilities to Disinformation and Data-Driven Propaganda", (Joint Research Centre (JRC) Technical Reports, European Commission, 2019), https://publications.jrc.ec.europa.eu/repository/bitstream/JRC116009/understanding_citizens_vulnerabilities_to_disinformation.pdf. On page 38 of the report it says: "They are designed to erode trust in mainstream media and institutions. Most of the content used to build these hostile narratives is not always objectively false. Much of it is not even classifiable as hate speech, but it is intended to reinforce tribalism, to polarize and divide, specifically designed to exploit social fractures, creating a distorted perception of reality by eroding the trust in media, institutions and eventually, democracy itself."

5 Donie O'Sullivan, "Lawmakers warn of 'deepfake' videos ahead of 2020 election", *CNN Business*, 28 January 2019, https://edition.cnn.com/2019/01/28/tech/deepfake-lawmakers/index.html; Donie O'Sullivan, When seeing is no longer believing - Inside the Pentagon's race against deepfake videos", *CNN Business*, https://edition.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/.

6 Marco Rubio website, "Rubio, Warner Express Concern Over Growing Threat Posed by Deepfakes", 2 October 2019, https://www.rubio.senate.gov/public/index.cfm/2019/10/rubio-warner-express-concern-over-growing-threat-posed-by-deepfakes.

7 Stephanie Kampf, Mark Kelley, "A new 'arms race': How the U.S. military is spending millions to fight fake images", *CBC News*, 18 November 2018, https://www.cbc.ca/news/technology/fighting-fake-images-military-1.4905775.

In short, in the disinformation arms race the capabilities available to malign actors are developing and proliferating at an unprecedented rate, while states and others developing defensive or protective countermeasures are struggling to keep pace.[8] As seen in the field of cybersecurity in the past, the emergence of new threats such as AI-supported disinformation campaigns will not be avoidable.

The remaining sections of this paper explore the next generation of AI-enabled information warfare, and consider the acknowledged, but as yet vague and abstract, threat of weaponising AI for malign influence campaigns. Section 2 discusses methods and prerequisites for utilisation of AI in modern information warfare. Section 3 reviews the state of the art of AI capabilities for generating and processing different types of information material, including the ability to identify, respond to and generate emotional response in human-machine and human-computer interaction. Section 4 draws on the previous sections and past observations to look forward into the next decade of AI-enabled information warfare and possible countermeasures to it, and finally section 5 recommends steps that NATO member states should take to mitigate this new type of threat.

2. AI IN INFORMATION WARFARE

In order to understand the capabilities of AI-supported disinformation campaigns, it is necessary to understand what can be achieved by the technology used. The true power of AI in information warfare derives from several factors: societies' reliance on social media; dependence on cyberspace as a trustworthy information resource; unlimited access to and ability to spread information rapidly through cyberspace; and human difficulties in reliably distinguishing between fake and genuine media, as well as a lack of authentication or validation capabilities online.

Malign influence campaigns in 2019 and before have involved a wide range of material being manipulated through different techniques and targeting different human modalities.

A. Methods

While there are many methods within the field of machine learning that can be used for AI applications, generative adversarial neural networks (GANs) became prominent for deepfakes during 2019. GANs utilise neural networks to optimise their output. In simple terms, a GAN is a couple of neural networks playing a game against each other (most commonly a zero-sum-game in terms of Game Theory). In the case of deepfakes, one neural network aims at building a deepfake from a set of input data, while the other aims at correctly distinguishing the deepfake from the original data.

8 Giles, Hartmann and Mustaffa, *The Role of Deepfakes*, 19–22.

Through this mechanism, the final output can be optimised with each "round" played. The method can be used both for the creation and alteration of media.

The artificial output produced becomes better over time and is also dependent on the required fidelity of the produced material. Typically, material of lower quality (image/video resolution or audio quality) is easier to fake, as there are fewer identifiable traits that must be learned. This has a direct effect on the amount of time needed for the training and hence on the time needed to produce a convincing deepfake.

From a technical perspective, there is a key difference between AI being used to create novel material and altering existing material. While the process involved varies slightly depending on the type of material being processed, the general concept remains similar. This allows an identification of the prerequisites needed, which is explored in the following subsection.

1) Creation

Currently, AI does not possess true creativity. Therefore, AI systems have problems generating unprecedented content, regardless of the type of output produced. However, what AI systems are particularly good at is learning correlations within data.

When producing novel content, AI systems tend to produce an average of the data used for training them. As an example: to produce a picture of an artificial woman, the AI will go through a database of images of women, extracting typical traits in those images in order to deliver an image containing the average of all identified traits. This is what most likely happened in the case of "Katie Jones". It also explains why she was identifiable as artificial through specific – yet minor – characteristics, such as blurred earrings of indefinable shape and colour. However, these artefacts of the AI processing can be avoided, either by manual post-processing of the generated output or by adjusting the AI accordingly.

Creating artificial content of a real, specific individual is slightly more complex and involves gathering training data on that particular individual. Publicly known individuals such as celebrities, politicians and major business leaders are therefore particularly at risk of being targets of AI-supported disinformation campaigns. Depending on the amount and quality of data available, creating artificial content may also involve application (and therewith learning) of general models, such as human-like movement patterns. This can then be used (to some extent) to compensate for a lack of sufficient data; however, it complicates the process and may be easier to identify as a fake.

2) Alteration

Compared to creation, alteration is somewhat more complex, as it involves more steps. In order, for example, to change a smile to a frown in a given picture, several steps are involved. First of all, the AI must understand which parts of a picture interact in order to be perceived as displaying a smile or a frown. These traits are universal to some extent, but may contain individual peculiarities. Hence, in order to be convincing, it is helpful to train the AI on the specific person whose image is to be altered. Having a model of how a smile (origin) and a frown (goal) look is the first step. The second step is to identify the areas that need to be altered. The third step is to perform the alteration and finally, the fourth step includes an adaptation to the overall image (such as light conditions, brightness and contrast). These steps are similar for video alterations.

Despite the fact that AI-supported alterations are somewhat more complex than generations, applications performing alterations already exist, as will be discussed further in section 3.

This kind of alteration should not be confused with simple editing, which continues to play an important role in disinformation. One prominent example from 2019 was a video of Ms. Nancy Pelosi, the US House Speaker and Democrat leader, which was altered to make her appear drunk and spread rapidly throughout social media.[9] This shows the potential of altered video material in misinformation campaigns generally. The case of Nancy Pelosi's altered video also showed some of the major concerns with social media. Despite the fact that the video gained 2 million views and had been shared 45.000 times within less than 36 hours,[10] Facebook confirmed that the video had been altered, but refused to take it down as "We don't have a policy that stipulates that the information you post on Facebook must be true."[11]

The Pelosi video was slowed down, making her speech appear slurred. Slowing down the replay rate of video and audio material is a very common task; most players have a function implemented for this purpose. Usually, the slower speed yields a notable change in the acoustics as well, resulting in a lower voice. In the case of Nancy Pelosi, the pitch had also been altered in order to compensate for this effect. While pitch alteration is not as common as change of the replay rate, it is still a task that requires little or no specific technical knowledge and is available on most audio and video processing applications. In a similar way, commonly available software for editing

9 Joan Donovan, Britt Paris, "Beware the Cheapfakes", Slate.com, 12 June 2019, https://slate.com/technology/2019/06/drunk-pelosi-deepfakes-cheapfakes-artificial-intelligence-disinformation.html.

10 Drew Harwell, "Faked Pelosi videos slowed to make her appear drunk, spread across social media", *The Washington Pos*t, 24 May 2019, https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/.

11 Drew Harwell, "Pelosi says altered videos show Facebook leaders were 'willing enablers' of Russian election interference", *The Washington Post*, 29 May 2019, https://www.washingtonpost.com/technology/2019/05/29/pelosi-says-altered-videos-show-facebook-leaders-were-willing-enablers-russian-election-interference/.

still images, audio files and text will continue to play a key role in malign influence campaigns alongside more advanced technologies.

B. Prerequisites

The arrival of big data processing methods, advances in computational power and parallel and distributed computing means that machine learning is no longer an exquisite technology available only to actors with enormous resources.[12]

As the multitude of deepfakes that arose during 2019 showed, the technology to produce deepfakes has become widely available. Some applications, such as "Zao"[13] and "FaceApp",[14] are available for download, while others provide online service platforms to create deepfakes.[15] While it is unlikely that these applications will be directly used in a disinformation campaign, the technology is being offered as a business product and thus is at a level that allows it to be used by software developers to create according applications and may hence also be used to develop applications for malicious use-cases. In section 4 we examine further how such a weaponised AI for disinformation campaigns may look today and how it is most likely to be enhanced in the near future.

3. AI APPLICABILITY

One particularity of the AI methods utilised in disinformation campaigns is that they may be applied to basically any material available. The reason for this lies in their pure and abstract nature: as long as there are specific patterns identifiable in a set of data, an AI construct will be able to identify, learn and reproduce the correlations between them. In the field of disinformation, however, the most relevant media are text, audio and video, and consequently the following section will give a brief overview of the current state of manipulation and creation technologies for each of these forms of material, including the ability of AI to identify and display human emotion within this material. This in turn will enable a better understanding of their future potential for disinformation campaigns.

A. Text

While most attention has been devoted to the disinformation potential of manipulated video, audio and still images, artificially generated text has been feasible for over a decade. In 2008 SCIgen, a scientific paper generator programmed by MIT students,

12 Samantha Cole, "This program makes it even easier to make deepfakes", vice.com, 19 August 2019, https://www.vice.com/en_us/article/kz4amx/fsgan-program-makes-it-even-easier-to-make-deepfakes; James Vincent, "AI deepfakes are now as simple as typing whatever you want your subject to say - A scarily simple way to create fake videos and misinformation", *The Verge Tech*, 10 June 2019, https://www.theverge.com/2019/6/10/18659432/deepfake-ai-fakes-tech-edit-video-by-typing-new-words.

13 Zao app, https://www.zaoapp.net/.

14 FaceApp, https://faceapp.com/app.

15 Deepfakes web β, https://deepfakesweb.com/.

managed to generate a research paper that was accepted by a conference (Computer Science and Software Engineering, CSSE, 2008, co-funded by IEEE) practising peer-review for publication.[16] While the purpose of SCIgen was to "auto-generate submissions to conferences that you suspect might have very low submission standards",[17] it also shows the extent to which even longer plausible texts may be generated automatically. A more recent release on the topic of artificial intelligence being used to produce artificial texts is OpenAI's GPT-2.[18] The text generator has already been identified as having the potential to produce propaganda or misinformation by extremist groups.[19] The implications for malign influence campaigns are multiple, including reducing or removing the reliance on humans to generate interactions, and thus solving the problem of scalability. Astroturfing, the practice of fraudulently generating messages designed to give the impression of widespread support for an idea, becomes vastly easier when it is not necessary to manually craft each message.

B. Audio

In the context of disinformation campaigns, the utility of audio material used to impersonate another individual is self-evident. During 2019 audio deepfakes, utilising the same technology used to create fake videos, were generated to impersonate the voices of CEOs by fraudsters in cybercrime cases.[20] One particular case described in more detail by *The Wall Street Journal* led to a loss of USD 243,000 through a fraudulent bank transfer.[21] The case shows the potential of the technology as well as the vulnerability presented by our reliance on the auditory identification of individuals. If they demonstrate target-specific knowledge, phone callers are often accepted as legitimate without having gone through a sufficient identification process; this is even more the case if the conversation is not about financial transfers but political opinions or personal statements.

In addition, the availability of speech synthesisers and their ability to generate artificial voices that sound human are on the rise. These systems are even capable of adding emotional prosody to the speech produced.[22] Like the example of text above, the clear implication is that disinformation campaigns will no longer be constrained by

16 The official "Herbert Schlangenman" blog, http://diehimmelistschoen.blogspot.com/.

17 SCIgen homepage at PDOS research group of MIT CSAIL, https://pdos.csail.mit.edu/archive/scigen/.

18 Irene Solaiman, Jack Clark, Miles Brundage, OpenAI Research Laboratory homepage and blog, "GPT-2: 1.5B Release", 5 November 2019, https://openai.com/blog/gpt-2-1-5b-release/.

19 Liam Tung, "OpenAI's 'dangerous' AI text generator is out: People find GPT-2's words 'convincing' -The problem is the largest-ever GPT-2 model can also be fine-tuned for propaganda by extremist groups.", ZDNet.com, 6 November 2019, https://www.zdnet.com/article/openais-dangerous-ai-text-generator-is-out-people-find-gpt-2s-words-convincing/.

20 Jesse Damiani, "A voice deepfake was used to scam a CEO out of $243,000", *Forbes*, 3 September 2019, https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/.

21 Catherine Stupp, "Fraudsters used AI to mimic CEO's voice in unusual cybercrime case - Scams using artificial intelligence are a new challenge for companies", *The Wall Street Journal*, 30 August 2019, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

22 Mark Schröder, "Emotional Speech Synthesis: A Review", Seventh European Conference on Speech Communication and Technology (Eurospeech 2001), Aalborg, Denmark.

numbers; but in this case an additional challenge that will be overcome is linguistic ability. In early 2019, one of the authors was the subject of a crude attempt at social engineering to assist a cyber exploit, where spear phishing victims received a phone call from an individual claiming to be the author's personal assistant and urging them to click on the link they had just received. Several of the victims were made suspicious by the caller's thick Russian accent – but once AI-generated synthesised voice capabilities are available, this will no longer be a limiting factor.[23]

Less relevant to the explicit context of disinformation campaigns, but a convincing demonstration of the capacities of AI in audio processing, is AIVA (Artificial Intelligence Virtual Artist): an AI system composing emotional soundtrack music.[24]

C. Video

Deepfake video came to widespread attention during the course of 2019, whether created for entertainment purposes or to raise the public awareness of deepfakes and their potential. Examples included videos where items or individuals were added to an existing clip, as well as existing videos being altered and new ones created.

Video alteration has involved the use of mouth models to adapt lip and face movement to make the speaker appear convincingly to be delivering the fake speech on the audio track. While the technology behind this involves many disciplines ranging from video processing, movement and biodynamic modelling to audio processing, the orchestration of tools generated within these research fields has led to the creation of applications usable by laymen that are fully capable of producing convincing footage.

D. Images

Image manipulation applications have become almost omnipresent on social media platforms, ranging from applications used to enhance self-portraits to those that add, delete or alter content within a picture. Newer applications utilising AI enhance this capability by creating photorealistic images from simplistic drawings[25] or artificial images based on machine learning algorithms ("Katie Jones"). Images may also be used to create video footage (see section 3. C).

E. Emotional Response Patterns

At the time of writing, the authors are not aware of instances of alteration of emotional states being displayed in images and videos. Nevertheless, this capability should be easily within reach. The Human-Computer Interaction (HCI) research community has

23 An overview of research projects in the field and their achievements can be viewed at http://emosamples.syntheticspeech.de/. The list is being maintained by Dr Felix Burkhardt, Director of Research at AudEERING GmbH (https://www.audeering.com/).

24 AIVA -The Artificial Intelligence composing emotional soundtrack music, https://www.aiva.ai/, sample tracks of AIVA can be found on YouTube: https://www.youtube.com/watch?v=gzGkC_o9hXI.

25 Nvidia AI Playground, Nvidia AI Research in Action, https://www.nvidia.com/en-us/research/ai-playground/.

devoted considerable effort to development both of systems capable of identifying, and virtual agents capable of displaying, human emotions. Videos simulating emotional reactions through facial movements are claimed to have been produced[26] from no more than a still image of a person and an audio clip.[27] The alteration of a video to include a simulated inappropriate emotional reaction could be a powerful tool to discredit public figures, especially as the changes made may be extremely subtle and hard to detect. A simple example could be a politician discussing a military operation that had claimed civilian victims, with his face altered to show indifference or even approval.

The HCI community has moved away from looking at what are known as the "Ekman basic emotions"[28] to concepts of more subtle emotional states and their transitions.[29] The research community has an excellent understanding of how emotions are being displayed and how to adapt systems to understand a specific users' hidden emotional cues.[30] However, with this knowledge, it is also able to reproduce footage displaying the subtle cues. Such alterations may even be difficult to identify for the individual being targeted, as many of these subtle emotional cues are a result of involuntary movements.[31]

4. THE NEXT DECADE

The weaponisation of AI for information warfare operations finds a natural home in cyberspace, an environment made up of pure digital data with no universal methods

26 Konstantinos Vougioukas, Stavros Petridis and Maja Pantic, "Realistic Speech-Driven Facial Animation with GANs", *International Journal of Computer Vision* - Special Issue on Generating Realistic Visual Data of Human Behavior, Springer, Online 13 October 2019, https://link.springer.com/article/10.1007/s11263-019-01251-8.

27 The video clips are available on YouTube: https://www.youtube.com/watch?v=NlNJKWPmmbk&feature=youtu.be.

28 A set of emotions that are cross-culturally recognisable, which were defined by Paul Ekman and his colleagues in a 1992 cross-cultural study. The emotions identified were: anger, distrust, fear, happiness, sadness and surprise. These have become generally accepted within the HCI research community as the "basic emotions".

29 Ingo Siegert, Kim Hartmann, Stefan Glüge and Andreas Wendemuth, "Modelling of Emotional Development within Human-Computer-Interaction", *Kognitive Systeme Journal* 2013, https://duepublico.uni-duisburg-essen.de/go/kognitivesysteme/2013/1/008; Kim Hartmann, Ingo Siegert, David Philippou-Hübner and Andreas Wendemuth, "Emotion detection in HCI: from speech features to emotion space." IFAC Proceedings 12th Symposium on Analysis, Design, and Evaluation of Human-Machine Systems, Volumes 46.15 (2013): 288–295, https://www.sciencedirect.com/journal/ifac-proceedings-volumes/vol/46/issue/15.

30 Simon Peter van Rysewyk and Matthijs Pontier, *Machine Medical Ethics*, (Springer, 2014).

31 Details on micro-expressions can be found through the Paul Ekman Group, a research group centred around Paul Ekman who also described the "Ekman Basic Emotions" (see footnote 28) and has produced various publications on the topic, https://www.paulekman.com/resources/micro-expressions/; The Facial Action Coding System (FACS) is being used by HCI researchers worldwide to identify emotional user responses during the course of HCI; Paul Ekman, Erika L. Rosenberg, "What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)", (Oxford University Press USA, 1997).

for authentication and validation of data. This is likely to have a number of direct effects on the conduct or execution of information warfare.

A. Command and Control

While in 2019 the process of generating a deepfake required human intervention, over the next decade this will become a far more automated process.

During the 2010s a successful disinformation campaign needed humans at every stage. The concept had to be developed, and the material needed to be designed, drafted, generated and spread through social media platforms. Dissemination required the utilisation of social media profiles, which needed to be created in advance of the campaign, often manually. These profiles needed to be serviced by humans in order to build social networks, generate followers and establish credibility. Hence, disinformation campaigns involved human labour and indeed formed a whole disinformation industry in countries like the Philippines, India and Russia.[32]

Examining automation already in use on social media platforms today does suggest it is unlikely that this heavy reliance on a human workforce will continue. The individuals involved are most commonly low-budget service providers operating with limited resources. The engagement of such operators in disinformation campaigns has several drawbacks, the most prominent ones being that they may accidentally (or, as in the case of the Internet Research Agency in St. Petersburg, Russia, deliberately) disclose details of their activities[33] – but in general they are less effective at operating covertly and are less efficient.

Platforms such as Instagram are already known for the presence of bot activities. Ingramer[34] provides Instagram bot services that take over users' account(s) and allow fully automated, simulated human behaviour on the platform. Ingramer even ensures that it cannot be tracked by Instagram through geo-location metadata; it performs actions such as like/follow/unfollow, direct messages, scheduled post, hashtagging, location and username targeting.

Similar bots and processes exist on most social media platforms. They have become relatively easy to develop, since most services/application providers of social media platforms allow developers to interact with the platform through developer application programming interfaces (APIs). These allow software developers to interact with the platform's application/service through their own code/applications.

32 Jonathan Corpus Ong, Jason Vincent A. Cabañes, *Politics and profit in the fake news factory – Four work models of political trolling in the Philippines*, (Riga: NATO STRATCOM COE, 2019), https://www.stratcomcoe.org/four-work-models-political-trolling-philippines.

33 EUvsDisinfo.eu, "Confessions of a pro-Kremlin troll", 26 April 2017, https://euvsdisinfo.eu/confessions-of-a-pro-kremlin-troll/.

34 Ingramer-Bots homepage, https://ingramer.com/.

Applications that address several APIs of different social media platforms are capable of controlling multiple accounts on multiple platforms. Such applications already exist and are available online. They are generally referred to as "social media management apps", and include examples such as Agorapulse,[35] Sprout Social[36] or Hootsuite.[37] The latter has a list of apps available that allow a connection of bots to Hootsuite Inbox.[38]

Since control panels for automated postings on social media are a mature, widely and cheaply available and broadly accepted technology, development of "command and control" panels for disinformation operations in hybrid warfare should be expected. Combining these with parallel developments in machine learning makes it likely that they will control AI agents (intelligent bots) capable of generating artificial content (semi-) automatically. The benefits are evident: a potentially unlimited number of accounts on multiple social media platforms that can be orchestrated by one individual, through one single application, spreading content generated by artificial intelligence pursuing a single and coordinated strategic goal.

B. Scalability

The technology used to produce deepfakes and other manipulated material is, at its core, nothing other than software. One goal for the weaponisation of AI for information warfare purposes in social media spaces is to automatically produce content that is coherent with the overall strategy of a disinformation campaign, but uses different means to display, share, and interact with the content produced. Due to the way social media works, this will heighten the trustworthiness of the content produced and ensure wide dissemination of the material.

As the scalability of software has been a major concern to the software engineering industry over the past years, especially with the shift towards "as a service" architectures, many concepts have been developed to allow an easy scaling of necessary software components. One of these concepts is microservice architectures, where each component of the software is a separate entity capable of operating on its own. This concept works very well with that of software agents. These entities (microservices) interact and respond to higher demands by automatically deploying several copies (instances) of themselves automatically through a so-called CI/CD (continuous integration/continuous deployment) pipeline. The CI/CD pipeline is part of "DevOps" (development operations) and the use of microservices with automated deployment is already industry standard for software engineers working on cloud architectures and other services needing to respond to changing demands.

35 Agorapulse: Social Media Management Software for Agencies and Teams, https://www.agorapulse.com/.
36 Sprout Social: Social Management Solution, https://sproutsocial.com/.
37 Hootsuite Social Media Tool – Schedule your Tweets, https://hootsuite.com/.
38 Hootsuite Apps – Bots – Apps that allow you to connect bots to Hootsuite Inbox, https://apps.hootsuite.com/categories/bots.

When designing a "command and control" panel as described above, it is reasonable to use a software engineering pattern that allows scalability. This will yield a more robust platform capable of producing high through- and output, where one panel is able to control hundreds of apparently independent accounts managed by software agents. This will ensure that performance limitations are negligible and allow a spontaneous adaptation to changing demands. The remaining limiting factor will be the control mechanisms installed by social media platforms, which are currently known to be insufficient.[39]

C. Automation

AI-assisted automation is very likely to be a major feature of the next decade of information warfare. This could apply in two distinct fields: automatically releasing disinformation following a coordinated overall strategy, and the automation of generating the disinformation. The latter task depends on acquisition of the data needed for the AI methods and their capability to generate content, preferably following a specific strategy (such as propaganda involving racism against a specific ethnicity). Automated release of already-available disinformation is easier to achieve, as it only requires scheduled access to the platforms targeted. From a technical perspective, this does not necessarily involve any artificial intelligence, although AI may be beneficial in order to create a more realistic illusion of human behaviour.

Automation could in the future also be used to generate instant responses to events. An intelligent information warfare campaign should be able to identify the rising interest in a relevant topic (such as the popularity of a specific individual or action) and generate a coordinated automatic response to leverage the interest. Response patterns could include producing counterarguments, fake news, "trolling" or cheap propaganda. In this context, the already existing abilities of AI systems to identify emotional states being displayed, to produce emotionally coloured responses, and to foresee their effects on humans will become of particular value. At present, all of these require a high number of user accounts sharing or promoting the produced material, which provides an obvious role for automation by more sophisticated means than the bots currently in use.

While the process of generating deepfakes is currently still being initiated manually, it should be expected that this too may soon be automated. However, producing still images to generate a profile picture of an artificial individual such as "Katie Jones" will still be far simpler than automatically generating a convincing deepfake video of an existing individual to deliver an automatically generated speech. It is likely that this type of activity will still involve a human workforce for the time being,

39 Sebastian Bay and Rolf Fredheim, *Falling behind: How social media companies are failing to combat inauthentic behaviour online*, (Riga: NATO STRATCOM COE, 2019), https://www.stratcomcoe.org/how-social-media-companies-are-failing-combat-inauthentic-behaviour-online.

until AI systems are capable of acting according to an abstract goal such as the one a disinformation campaign may have.

As described in section 3. A, the production of shorter texts with the aid of AI when given a set of keywords is already reality. Having bots active on social media platforms that post these artificially generated texts is not a challenge. Even today, social media users such as influencers manage their account(s) through applications that allow them to schedule pre-defined posts or to generate posts out of a set of texts, hashtags and pictures. It is likely that similar techniques, enhanced through machine learning, will be deployed in information warfare in the near future, augmenting troll factories and botnets.

D. Countermeasures

It is important to understand that the processes described in this paper do not depend on future or emerging technologies. Each of the capabilities required is already at an advanced stage, and the respective research fields have developed these technologies for specific and multi-modal systems over the past years and, in some cases, decades. The techniques are ready for use in legitimate civilian applications, and in many instances are already known to be being weaponised by malicious actors. This is a matter for real and urgent alarm, since AI-supported disinformation campaigns have the potential to impose the largest threat to democracy and society seen so far, targeting not only public opinion but the nature of belief and trust, which constitute pillars of democratic societies.

In common with other new technologies, it is very unlikely that weaponisation can be prevented. Instead, methods to authenticate and distinguish original from manipulated material on a mass scale and in real time are urgently required. The particular problem with identifying manipulated material lies within the methods used to generate this material: GANs, as described above, if sufficiently well-trained, will yield an outcome that is difficult to distinguish from genuine media – not just for the human observer, but also for machines.

An alternative approach could be certification of genuine material. In the same way as the internet as a whole was designed to be insecure, meaning that secure applications and processes needed to be developed separately, so in an information space that is generally untrusted, additional measures could be necessary to stimulate trust. Possible technologies for doing so could include digital watermarks (requiring the involvement of manufacturers to include the technology in recording devices), or software signature processes, as known in e-mail communication. In both cases, however, this kind of approach would only be of use for a small subset of the total amount of information in circulation. The disinformation industry relies heavily

on propaganda being spread through social media. The material being spread does not necessarily have to appear official, as long as it is convincingly real, or at least plausible, and provides an explanation of how or why someone got access to the record. At the same time, the widespread consumption of this type of material has contributed to the public becoming accustomed to low-quality material originating from doubtful sources and claiming to show the real truth to a story. This eases the task of malicious actors intending to spread disinformation.

A third approach that requires further investigation is that of using distributed knowledge to validate material being circulated. The idea is to use the knowledge of several individuals, sensors and general information, combined to reason the validity of the material being displayed. This combined knowledge could include verification by known witnesses, physical phenomenon validity checks (e.g. light effects or interactions between the environment and objects in videos), surveillance monitoring data and background information checks (such as validation of the caller ID in telephone calls through the service provider or specific knowledge of the location being filmed).[40] Notable results may be derived from research into swarm intelligence, a subfield of artificial intelligence.

5. OUTLOOK

Information warfare lies at the intersection of several well-established trends that will combine to pose severe challenges to nations and societies in the short and medium term. These are:

- The continuing progress of hyperconnectivity, reducing the perceptibility of dividing lines between online and real life;[41]
- Reduced restraint by actors hostile to liberal democracies, as they are emboldened by the apparent lack of deterrent measures available to their targets;
- Further erosion of trust, and of the notion of independent and verifiable truth;[42]
- Finally, as detailed in this paper, the rapid and accelerating pace of change in technologies that facilitate or enable malign influence campaigns.

40 Jack Corrigan, "DARPA Is Taking On the Deepfake Problem", NextGov.com, 6 August 2019, https://www.nextgov.com/emerging-tech/2019/08/darpa-taking-deepfake-problem/158980/, "A comprehensive suite of semantic inconsistency detectors would dramatically increase the burden on media falsifiers, requiring the creators of falsified media to get every semantic detail correct, while defenders only need to find one, or a very few, inconsistencies,"; Derek B. Johnson, "The semantics of disinformation", Defensesystems.com, 26 August 2019, https://defensesystems.com/articles/2019/08/26/darpa-disinformation-semantics-johnson.aspx.

41 Kim Hartmann and Keir Giles, "Shifting the core: How emergent technology transforms information security challenges", *Datenschutz und Datensicherheit (DuD)*, Springer Journal, 14 June 2017, https://link.springer.com/article/10.1007/s11623-017-0807-y.

42 Giles, Hartmann and Mustaffa, *The Role of Deepfakes*.

Of these parallel but interdependent phenomena, perhaps the most straightforward to prepare for is the impact of technologies. Unlike the other trends, this is both relatively predictable and has a set of clearly identifiable countermeasures.

These could include:

- Exploring methods of technical authentication of digital material;[43]
- Content provenance through digital signatures;[44]
- Considering further applications of digital signatures;[45]
- Continuing efforts to restore trust in independent media and journalism;
- Inducing social media platforms to enhance the detection of fakes and to install means to allow users to evaluate the reliability of content;
- Ensuring the availability of national or supranational authorities to which civilians can report instances of malign influence campaigns;
- Following the example of Singapore,[46] forcing social media platforms to mark fake or false content (including any repost or shared post of the initial material).

Each of the new technologies detailed in this paper will have an impact on information warfare; but it need not be a transformative one. As with other previous technological developments, delivery of disinformation may be effected in a different manner, but the fundamental nature of deception remains unchanged. As such, the basic ingredients of countering it follow the same pattern as in previous decades and indeed centuries. This is because while disinformation techniques and technologies change, one factor that remains constant is the human susceptibilities they exploit.

It follows that alongside the technical recommendations above, individual states should undertake clear and honest assessments of their own publics' susceptibility to malign influence campaigns. Metrics to quantify and understand this susceptibility are as urgently needed as metrics to assess the success or failure of disinformation campaigns, and are essential both to preparing countermeasures and to fostering societal awareness of the threat.

A wide range of factors determine how susceptible a community is to disinformation: access to and engagement in social media, media uptake and trustworthiness, age,

43 Antonio García Martínez, "The blockchain solution to our deepfake problems", *Wired Magazine*, 26 March 2018, https://www.wired.com/story/the-blockchain-solution-to-our-deepfake-problems/.

44 National Academics of Sciences, Engineering, and Medicine, Chapter 6 "Deepfakes" in *Implications of Artificial Intelligence for Cybersecurity: Proceedings of a Workshop*, (The National Academies Press, Washington D.C. 2019), https://doi.org/10.17226/25488.

45 Kalev Leetaru , "Why digital signatures won't prevent deep fakes but will help repressive governments", *Forbes*, 9 September 2018, https://www.forbes.com/sites/kalevleetaru/2018/09/09/why-digital-signatures-wont-prevent-deep-fakes-but-will-help-repressive-governments/.

46 Singapore Legal Advice, "Singapore Fake News Laws: Guide to POFMA (Protection from Online Falsehoods and Manipulation Act)", 2 January 2020, https://singaporelegaladvice.com/law-articles/singapore-fake-news-protection-online-falsehoods-manipulation/.

technical education level, trust and understanding of democratic values, as well as trust in national leaders. The latter point places an obligation on leadership figures in Western liberal democracies to understand their own contribution to societal cohesion and common defence. Trust in leaders and institutions is a foundation stone of democratic systems; and an erosion of this trust through flagrant disregard for honesty and probity while in power creates a power vacuum which can and will be exploited by adversaries.

In keeping with all of this, a recommendation that remains common to all counter-disinformation efforts is raising public awareness: of the threat, of its methods, and of the indicators and warnings that an individual or group is being subjected to a malign influence campaign – critically, regardless of whether this campaign is mounted by foreign adversaries or domestic political actors manipulating society for their own ends. To this long-standing recommendation should now be added audience education on the nature and capabilities of the next generation of AI-enabled disinformation technologies. Proper preparation and investment in threat literacy among target audiences, started now, will have a substantial impact in mitigating the potential dangers of information warfare in the next, even more complicated decade.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Defenders Disrupting Adversaries: Framework, Dataset, and Case Studies of Disruptive Counter-Cyber Operations

Jason Healey
Senior Research Scholar
School of International and
Public Affairs
Columbia University
New York, NY, United States
jh3639@sipa.columbia.edu

Neil Jenkins
Chief Analytic Officer
Cyber Threat Alliance
Arlington, VA, United States
neiljenkins@cyberthreatalliance.org

JD Work
Bren Chair, Cyber Conflict and Security
US Marine Corps University
Quantico, VA, United States
JW3646@columbia.edu

Abstract: Over the past two decades, there have been numerous defensive operations to disrupt malicious cyber activity by hacktivists, criminals, and nation-state actors. Disruption operations seek to affect the adversary's decision-making processes and impose additional costs. Such operations include a wide range of actions, from releasing indicators of compromise and naming-and-shaming, to botnet and infrastructure takedowns, to indictments and sanctions, and may be conducted outside of the defender's own network with the intent to interrupt adversary cyber offense and espionage. The United States Department of Defense recently released a new strategy that calls for "persistent engagement" with malicious cyber actors, suggesting many more disruption operations to come.

In this paper, we describe a framework for categorizing disruption operations and their effects – along with detailed descriptions for several of these case studies coded to the framework – so that researchers and practitioners can measure their impact using a common terminology. We also provide a unique dataset of over 100 cases of defensive operational disruption over the last 30 years, from 1987 through 2019.

We believe that providing a more complete vocabulary for disruptive operations will give analysts and researchers a better opportunity to compare the different types and effects of various disruption operations. Ideally, this will then provide defenders with the information they need to conduct disruption operations at greatest scale, least cost, and with the lowest chance of escalation.

Keywords: *offensive cyber, counter-cyber, takedown, disruption*

1. INTRODUCTION

The United States military has reoriented its role in order to emphasize a "persistent presence" to "intercept and halt cyber threats" with the hope of countering "malicious cyber activity in day-to-day competition".[1] Through persistent engagement, the DoD will employ defensive cyber operations to disrupt adversaries' operations directly and impose friction so they will be forced to spend more resources on defense, rather than offense.[2]

However, there is no public methodology that can measure the effectiveness of such disruptive operations. Without a measurement methodology, analysts cannot reliably assess the success of this policy or compare the effectiveness of different kinds of disruptive operations. Building upon earlier work by Healey and Jenkins in measuring the effects of persistent engagement, this study builds toward understanding the real-world impacts of such operations.[3] This paper begins by describing an analytical framework for assessing disruption operations, which is followed by an assessment of five cases using the framework, including a unique dataset of 100+ such cases. A concluding section summarizes the insights, future research, and conclusions.

1 Department of Defense. *Cyber Strategy*. 18 September 2018.
2 Jason Healey, "The Implications of Persistent (and Permanent) Engagement in Cyberspace," *Journal of Cybersecurity*. 5, no. 1 (2019).
3 Jason Healey, Neil Jenkins. "Rough-and-Ready: A Policy Framework to Determine if Cyber Deterrence is Working or Failing." *11th International Conference on Cyber Conflict: Silent Battle. Tallinn, Estonia. 28-31 May 2019*.

Though these are still early steps, our goal is to encourage transparency and repeatability to better characterize and understand the scope and range of disruptive counter-cyber operations. We explore the factors that lead to the "most effective" disruption outcomes, although a more complete assessment is out of the scope of this paper. In general, we anticipate that disruptive actions that are more active, more collaborative, more frequent, and more intrusive will have greater impact. But we recognize that mere attrition is not the only measure of effect, as some disruptive actions will likely offer more decisive effect at some substantive threshold, or within particularly operationally relevant timeframes. We anticipate that the elements contributing to successful disruption outcomes will vary across differing situations, and that while a simplified generalization of best choices is not likely possible, there are specific most-effective approaches for a given type of disruptive activity.

2. ANALYTICAL FRAMEWORK

Disruptive counter-cyber operations are positive steps for defeating a specific cyber adversary, usually taken by defenders in response to a specific attack or campaign, and they often directly disrupt an adversary's technology; the main action is typically either outside of the defender's own network or based on specific intelligence about how that adversary operates. This is only a general description, as each element of that description contains important exceptions, so we will examine each part individually:

1. **Positive steps to defeat a *specific* cyber adversary**, usually but not always conducted online. It would not include best-practice defensive measures, such as patching computers, unless specifically intended to defeat a particular adversary that is known or suspected to be targeting that vulnerability. Disruptive operations are generally marked by *active contention* with an adversary.
2. **Usually taken by a defender**, such as a government, cybersecurity, or technology company, or the victim of an attack. There are rare exceptions, such as examples of so-called red-on-red operations where two maliciously motivated actors contest control of infrastructure for their own objectives that remain at odds with the victim's interests.
3. **Taken in response to a specific cyber attack or campaign** to disrupt an adversary's ability to continue ongoing action. This distinguishes it from offensive cyber effects operations (which may come before, during, or after a campaign and serve different purposes), pure retaliation (which is meant to punish for past, not disrupt ongoing, behavior), or deterrence-by-punishment (which is intended primarily to punish an adversary to change their decision calculus). This framework is only, for now, interested in *disrupting* cyber

activities (such as disruptive attacks or intrusions) and not *influence* or *information* operations. We include some actions, such as law-enforcement indictments, in this framework, which may take place well after a campaign. However, these share enough other characteristics with other disruptive operations to be usefully included.

4. Often **directly disrupt an adversary's technology** and typically the **main action is outside of the defender's own network** or **based on specific intelligence** about how the adversary operates. A botnet takedown disrupts technology outside the network of most defenders, while cybersecurity companies and infrastructure sectors share, routinely and at massive scale, their insights of adversary groups to block their efforts on defenders' internal networks.

We evaluate such disruptive operations through a framework of multiple factors related to execution, approach, impact, and adversaries. This framework is neither a formal taxonomy nor has it matured through extended use by analysts; rather it is intended as a first draft of an analytical tool.

A. Dependent Variable: Effect and Duration of Disruption

The effectiveness of disruptive operations is the dependent variable, the thing we want to explain. It can be assessed in at least two ways, a simple description of the impact as well as an estimate of how long it takes the adversary to return to initial operating capability (able to conduct some limited operations) and return to full operating capability (approaching the full range of the adversary's previous activity). These measures of effect and duration overlap; and with use, it may be obvious which of these two is most useful. As that is not yet clear, both are included here.

Effect can be described by a simple three-point scale:

- *Minor*: Slight impact to adversary operations;
- *Significant*: Intermediate impact;
- *Decisive*: Substantive impact.

Duration can be hard to measure, so is simplified to a four-point scale:

- Days to weeks;
- Weeks to months;
- Months to years;
- Never.

A disruption might be so massive that the adversary group disbands. In these cases, the mission, personnel, tools, or infrastructure may be handed off to other groups associated with a particular nation or group, which can confound this assessment.

The other elements of the framework categorize the independent variables, those which will be studied for the impact on the effectiveness of this dependent variable of disruption.

B. Independent Variables

1) Type of Disruption

Technical measures to disrupt adversaries cover a wide spectrum and can usefully be categorized in many ways. For example, disruptive operations can be categorized by the *functional object* at which they are targeted:

- Systems and infrastructure in blue space (that is, owned or operated by the defenders);
- Systems and infrastructure in gray space (owned by neither defenders nor adversary);
- Systems and infrastructure in red space (owned or operated by the adversary);
- Command-and-control (C2) capabilities;
- Adversary personnel;
- Adversary organizations;
- Adversary leadership.

Another categorization is by the *action*, from relatively passive to far more active measures:[4]

- Sinkhole traffic;
- Share threat intelligence with closed trust group (multiple security actors);
- Publicly disclose indicators of compromise;
- Publicly release adversary toolset;
- Publish comprehensive report on malicious cyber activity and mitigations;
- Build protections for security products based on observed indicators of compromise and behaviors (single actor);
- Synchronize the deployment of protections (multiple actors);
- Coordinate vulnerability patching or other protections;
- Disrupt criminal channels for distribution or monetization;
- Force uninstall/deletion/takeover of malware;
- Seize control of adversary C2 nodes or network;

4 The authors have conducted an initial cross-linking of these actions against the Lockheed Martin's Cyber Kill Chain, though it is not included here for brevity. While useful, frameworks like the kill chain have some limitations as they are private-sector focused and lack a feedback loop through which to include the impact of disruptive actions.

- Seize domains used by adversaries;
- Counter-offensive operations to directly target intermediate infrastructure;
- Counter-offensive operations to disrupt attackers' home networks;
- Seizure or kinetic destruction of servers or infrastructure.

Disruptive actions also can be directed not at an adversary's technical infrastructure but their *decision making*:

- Publicly disclose the identities or organizational affiliation of the adversaries;
- Publicly disclose the nation responsible;
- Diplomatic démarche;
- Law-enforcement indictment and prosecution;
- Influence operations against individuals, organizations, or leadership;
- Deception operations;
- Economic sanctions;
- Military options (kinetic or cyber) to coerce adversary to desist.

These actions reflect a range of defensive cyber operations measures, response actions and other counter-cyber operations options, and full offensive employment approaches. These are commonly defined within the US Department of Defense and allied doctrine, which in turn is adopted directly or through influence of common practice by other actors across the environment.[5] The decision to select one set of options versus another is highly case-specific. This decision is influenced by the identity and available authorities of the disrupting actor, available technical capacity and talent, target-specific vulnerabilities and operational security failures, adversary organizational and process considerations that may be variably exploited, as well as temporal considerations.

2) Frequency of Disruptive Activity
Disruptive operations can take place with different frequencies:

- *One-off*: Disruptive activity is only conducted once;
- *Periodic*: Related set of disruptive activities taking place occasionally over time;
- *Sustained*: Related set of disruptive activities taking place frequently and in a coordinated manner.

3) Potential Reasons for Delay in Returning to Operations
Adversaries may not return to full operating capability for reasons only loosely related to the disruptive action. Accordingly, any analytical framework must include some

5 Department of Defense. *Cyberspace Operations*. Joint Publication 3–12. 8 June 2018.

way to include such assessments lest defenders misunderstand the actual impact of their operations. These factors include the following:

- *Technical*, for example from having attack infrastructure burned;
- *Behavioral*, such as if adversaries shift to a different, less fruitful, target set;
- *Bureaucratic*, perhaps from a re-organization once certain adversary teams were publicly called out;
- *Political*, for example if adversary leadership shift operations to favor other domestic interest groups or cut down on operations seemingly out of their control or linked to corruption;
- *Geopolitical*, if an adversary fears backlash for operating against another nation.

4) Geopolitical Context of Disruption
Analysts must also distinguish the geopolitical context of the disruptive operation, which will often have significant explanatory power as other elements:

- *Peace*: Lack of any significant military or diplomatic confrontation;
- *Tension*: Increase of military or diplomatic confrontation but unlikely to escalate into war without significant additional degradation;
- *Crisis*: Significant, acute military or diplomatic confrontation, especially with a chance of war or substantial national interests at stake;
- *War*: Active and routine military operations by the participants.

5) Type of Adversary
Lastly, the framework must distinguish both what kind of organization is conducting the disruption and what kind is being disrupted:

- Disrupt*ed* actor (adversary): state/non-state, criminal or geopolitical aims, relative cyber maturity, relations with other adversary groups, etc.;
- Disrupt*ing* actor: state, major technology company, geopolitical, coalition of states, coalition of technology groups, public–private sector partnership (PPP), etc.

3. CASE STUDIES OF DISRUPTIVE COUNTER-CYBER OPERATIONS

Multiple incidents provide ample fodder for case analysis to use this framework. This section first introduces our dataset (see Table 1 below), consisting of over 100 cases of defensive operational disruption over 30 years, from 1987 through 2019, and then

explores five case studies. These cases were selected based on industry intelligence reporting and information security literature, in which specific actions were noted to have had impact on adversary evolution, changing capabilities and intentions, or future operational planning for later disruption actions. While the influence of these cases can be traced in multiple intelligence and operational contexts, no prior effort to systematically assemble, document, and assess the corpus in total could be identified.

The limitations of space preclude comprehensive examination of each incident. Each case arose within the context of a specific threat exploiting discrete vulnerabilities to deliberate effect, often reported on over months or years in a body of work that alone may fill entire volumes. However, several cases are especially relevant as illustrative examples of the proposed assessment framework.

Eviction of CodeRed Worm: "White worm" inoculated vulnerable systems by unknown and red-on-red actors (rows 4-6 in dataset)

The widespread propagation of the CodeRed worm across Microsoft Internet Information Services (IIS) servers vulnerable to CVE-2001-0500 in July 2001 was a formative event for many cybersecurity professionals, and was among the first incidents in which political motivations were widely considered due to geographic references left in the malware itself. The incident drove substantial efforts toward information sharing, collaborative defense, and crisis management practices that remain fundamental to the industry. However, a perceived lack of effective government response further drove early vigilante efforts to degrade the effectiveness of adversary action, resulting in the release of one of the earliest examples of "white worm" deployment, in which payloads intended to inoculate vulnerable systems were released into the wild by unknown actors – without the consent of system owners. Ultimately, the CodeGreen "vaccination" campaign would itself serve as a model for further adversary abuse where other adversary actors sought to deliver their own wormable payloads exploiting the same vulnerabilities – evicting CodeRed infections but also delivering control of these systems to other operators with hostile intent in one of the earliest documented adversary on adversary (red-on-red) campaigns.[6] The case, despite its age and some complexity across multiple incident phases, remains significant, as both criminal and advanced persistent threat group predation on other vulnerable bad actors continues to surface as an ongoing feature of the contemporary cyber environment.

6 David Moore, Colleen Shannon, and K. Claffy. "Code-Red: A Case Study on the Spread and Victims of an Internet Worm." *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurement (IMW). Marseille France. November 2002.*; Nicholas Weaver, Vern Paxson, Stuart Staniford, and Robert Cunningham. "A Taxonomy of Computer Worms," *WORM '03: Proceedings of the 2003 ACM Workshop on Rapid Malcode, Washington, DC, USA, 27 October 2003*, https://doi.org/10.1145/948187.948190.

- Effect and duration of disruption: Minor; while the white worms reduced the pool of some vulnerable systems, the scale of overall vulnerability still resulted in substantial adversary freedom of action;
- Type of disruption: Forced uninstall/deletion/takeover of malware;
- Frequency of disruption: Sustained (for inoculated systems);
- Potential reasons for delay: Technical;
- Geopolitical context: N/A;
- Type of adversaries: Criminal, Unknown and red-on-red.

Conficker Disruption: long-term counter-malware campaign (row 21 in dataset)

The sustained, multi-stakeholder effort required to disrupt widespread infections of the serial version of the Conficker malware family provides an instructive case of effective disruption operations involving large scale, rapidly evolving threats. The years-long efforts of a group of quiet professionals, often working with only limited government support and against a backdrop of serious litigation, policy, and financial risks, is the stuff of legends among the infosec community.[7] They employed counter-measures – especially sink-holing operations – to halt propagation and defeat hostile administration of compromised victims through seizing domain registration, which was complicated by the wide, algorithmically derived namespace used for malware C2. The ultimate resolution of this case is intertwined in subsequent exploitation of common vulnerabilities, to the point that the arrests of several Conficker operators in the Ukraine passed largely unnoticed.[8] The ability of the disruption operators to generate and maintain pressure on the botnet severely limited the adversary's ability to leverage any utility of what was an innovative and even surprising design.[9]

- Effect and duration of disruption: Significant, weeks to months;
- Type of disruption: Botnet takedown with multiple active and passive measures, targeting technical infrastructure and actions in blue, gray, and red space;[10]
- Frequency of disruption: Sustained;
- Potential reasons for delay: Technical;
- Geopolitical context: N/A;
- Type of adversaries: Criminal, PPP.

7 Mark Bowden, *Worm: The First Digital World War* (London: Atlantic Books, 2012).

8 Brian Krebs, "$72M Scareware Ring Used Conficker Worm," *Krebs on Security*, June 2011, https://krebsonsecurity.com/2011/06/72m-scareware-ring-used-conficker-worm/#more-10417.

9 Dave Piscitello, "Conficker Summary and Review," *ICANN*, 7 May 2010, https://www.icann.org/en/system/files/files/conficker-summary-review-07may10-en.pdf.

10 The formulation of blue, gray, and red network space is taken from current USG operational thinking, which makes key distinctions between friendly (blue) and adversary systems and networks (red), as those which are effectively uncontrolled (gray).

GameOverZeus Takedown: heavily coordinated takedown of a botnet by a public–private partnership (row 56 in dataset)

Operation Tovar, the takedown against the GameOverZeus botnet in June 2014, was more technically complex than any preceding it, due to the resilient peer-to-peer C2 architecture, itself evolved under earlier and continuing administrative, technical, and law-enforcement pressures. The botnet was disrupted through cryptanalytic attack under judicial authorities, exploiting weaknesses in C2 protocol to contest adversary control of infected bots through forged commands issued via disruptive nodes introduced into peer-to-peer exchange. As a result of worldwide law-enforcement actions coordinated with technical action, the adversary was unable to resist loss of infrastructure.[11] However, this action may have represented the high water mark for law-enforcement-led, public-private partnerships to counter malicious infrastructure, as it has been suggested that subsequent takedown efforts have been increasingly less effective over time.[12]

- Effect and duration of disruption: Decisive, months to years;
- Type of disruption: Botnet takedown with multiple parallel active measures against technical infrastructure, owned by adversary;
- Frequency of disruption: One-off;
- Potential reasons for delay: Technical;
- Geopolitical context: N/A;
- Type of adversaries: Criminal, PPP.

China-Related Disclosures: public disclosure of cyber espionage (rows 45, 55, 65, 82, and 91 in dataset)

Public attribution linking Chinese operators to ongoing intrusion campaigns remains a vital tool for many states seeking to challenge the undesirable behavior of competitors in the court of public opinion, intended to impose political costs on adversary actors as well as their sponsors and leaders.[13] There is some evidence to suggest that the sequential impact of mere disclosure may be attenuated when hostile services are repeatedly accused – whether through *name and shame* as an influence tactic, or even indictments under judicial process. The initial disclosures linking the APT1 / Comment Crew intrusion set to the operations of a specific People's Liberation Army unit had

11 Europol, "International Action Against 'Gameover Zeus' Botnet And 'Cryptolocker' Ransomware," News release, (2 June 2014).; Symantec, "International Takedown Wounds Gameover Zeus Cybercrime Network," News release, (2 June 2014).; Brian Krebs, "'Operation Tovar' Targets 'Gameover' ZeuS Botnet, CryptoLocker Scourge," *Krebs on Security*, (2 June 2014), https://krebsonsecurity.com/2014/06/operation-tovar-targets-gameover-zeus-botnet-cryptolocker-scourge/.

12 Brandon Levene, "Crimeware in the Modern Era: A Cost We Cannot Ignore," *Chronicle*, 5 September 2019.

13 Florian J. Egloff and Andreas Wenger, "Public Attribution of Cyber Incidents," *Center for Security Studies, ETH Zurich. May 2019*.

substantial diplomatic impact.[14] It is likely that the multi-year reverberations of this action were a contributory factor to the 2015 agreement between Xi and Obama to prohibit further economic espionage, wherein both sides agreed that "neither country's government will conduct or knowingly support cyber-enabled theft of intellectual property, including trade secrets or other confidential business information, with the intent of providing competitive advantages to companies or commercial sectors."[15]

The sequential disclosures of multiple intrusion sets attributed to the Ministry of State Security – including APT3 / GOTHIC PANDA / UPS TEAM, APT10 / STONE PANDA / MenuPass / POTASSIUM, and APT17 / AURORA PANDA / DOGFISH – each challenged the earlier narrative of diplomatic agreement, in which China was seen as a reformed actor, adhering however loosely to the spirit of the negotiation.[16] Industry reporting on these intrusion sets' victims, accesses, and action objectives was matched by an unknown third party disclosure offering substantial attribution detail, followed by Department of Justice indictments.[17] Open questions remain, however, as to whether this strategic impact translates to operational disruption effect.

- Effect and duration of disruption: Unknown;
- Type of disruption: Disclosure;
- Frequency of disruption: Periodic;
- Potential reasons for delay: Bureaucratic (intelligence gain / loss considerations, diplomatic concerns);
- Geopolitical context: Tension (great power competition);
- Type of adversaries: State intelligence, State intelligence / Law Enforcement, unknown actor(s).

Joanap Takedown: government takedown of botnet (row 94 in the dataset)

The Joanap botnet was a component of the infrastructure used by the DPRK-attributed HIDDEN COBRA / LAZARUS intrusion set for reconnaissance, staging, and

14 FireEye, "APT1," 19 February 2013, https://www.fireeye.com/content/dam/fireeye-www/services/pdfs/mandiant-apt1-report.pdf.

15 White House, "FACT SHEET: President Xi Jinping's State Visit to the United States," 25 September 2015, https://obamawhitehouse.archives.gov/the-press-office/2015/09/25/fact-sheet-president-xi-jinpings-state-visit-united-states.

16 FireEye, "Red Line Drawn: China Recalculates Its Use of Cyber Espionage," 21 June 2016; Robert Farley, "Did the Obama-Xi Cyber Agreement Work?" *The Diplomat*, 11 August 2018.; Herb Lin, "What the National Counterintelligence and Security Center Really Said About Chinese Economic Espionage," *Lawfare*, 31 July 2018.

17 DOJ, "U.S. Charges Three Chinese Hackers Who Work at Internet Security Firm for Hacking Three Corporations for Commercial Advantage," 27 November 2017. https://www.justice.gov/opa/pr/us-charges-three-chinese-hackers-who-work-internet-security-firm-hacking-three-corporations; Cristiana Brafman Kittner and Ben Read. "Red Line Redrawn: China APTs Resurface," *FireEye Cyber Defense Summit. Washington, DC. 1-4 October 2018*.; DOJ, "Two Chinese Hackers Associated with the Ministry of State Security Charged with Global Computer Intrusion Campaigns Targeting Intellectual Property and Confidential Business Information," 20 December 2018. https://www.justice.gov/opa/pr/two-chinese-hackers-associated-ministry-state-security-charged-global-computer-intrusion

distributed denial of attack (DDOS) actions, leveraging previously infected victim systems at scale for multiple global operations since at least 2009.[18] The infrastructure was reportedly quite aged at the time of the January 2019 takedown operation by the Department of Justice and Air Force Office of Special Investigations, and considered "not that interesting" by industry researchers.[19] Still, the takedown action against this legacy infrastructure precluded adversary options to later revive it, especially under the continuing pressure of other ongoing countering options intended to deny and degrade North Korea's cyber operations posture. And even if the adversary had not intended to return this legacy inventory of compromised bots to active use, the takedown effort to disrupt potential hostile use of these bots may be viewed in analogy to removing unexploded ordnance. While removing unexploded ordnance may not be considered the most impactful mechanism in a contest with an adversary offensive program, such actions have undeniable value for the stability of the global cyberspace ecosystem as a whole.

- Effect and duration of disruption: Decisive, months to years;
- Type of disruption: Botnet takedown;
- Frequency of disruption: One-off;
- Potential reasons for delay: Technical and bureaucratic (may not have been worth devoting resources to building an obsolete network);
- Geopolitical context: Tension;
- Type of adversaries: State intelligence, law enforcement.

4. DATASET OF OPERATIONAL DISRUPTION

A full coding of all 100+ cases in this framework is outside the scope of the current paper. Rather we have used a simplified coding, starting with a common name for the operation or disrupted group and the approximate date of the operation. The third column codes the motivation of the disrupted adversary, whether criminal, hacktivist, espionage, or strategic attack. Motivation is coded based on contemporaneous reporting assessment by the security researchers, commercial intelligence firms, or government actors involved in the action. While this potentially omits later understanding of complex motivations developed through deeper historical analysis, it does capture the then-dominant consensus views and therefore the key influences involved in disruption actions at the time when these decisions were taken.

In a few cases, the disruption was not related to targeting an adversary but had another purpose, such as inoculation, essentially intruding into others' vulnerable devices to pre-emptively patch them against the truly malicious. Those cases are coded as

18 DHS CISA, "HIDDEN COBRA – Joanap Backdoor Trojan and Brambul Server Message Block Worm," 29 May 2018, https://www.us-cert.gov/ncas/alerts/TA18-149A.

19 Amit Serper. ""Hmm wait. This is ancient stuff, from like... errr... Almost 7 years ago. OTH, not that interesting?" 4 June 2018, Twitter, https://twitter.com/0xAmit/status/1003742265762811905.

vulnerability reduction.[20] The last column codes the actor conducting the disruption: industry, government, or public-private partnerships. A small number of cases are red-on-red incidents between malicious adversary operators. Those coded as government can be further specified as intelligence, military, law enforcement (LE), or national Computer Emergency Response Teams (CERT).[21] However, to date, we have only documented LE cases. In some LE cases, the originating investigations may have been enabled by unacknowledged industry support, and government intelligence services may play an unacknowledged role in many other cases in ways that have not been publicly documented to date. No unilateral CERT actions have as yet been identified in these cases, likely due to the collaborative nature of these organizations' work processes in coordinating action on private sector networks, inherently involving public-private partnership. Despite this, some unilateral responses may be contemplated and the option to recognize these edge cases is preserved. The dataset deliberately excludes actions to counter hostile influence operations and other coordinated inauthentic activity conducted through cyber platforms, as we are focusing for now on "hard" offensive cyber interactions.

The dataset is skewed toward open-source reporting, as industry and law enforcement often disclose operations for public relations value.[22] Longer-term exploitation of targeted adversary infrastructure through counter-cyber network exploitation (CCNE) operations is likely underrepresented, including in LE cases where employment of active network investigative techniques may have preceded takedown actions.[23] The use of such techniques has been documented in multiple contexts, but, due in no

20 Nicholas Weaver, Vern Paxson, Stuart Staniford, and Robert Cunningham. "Large Scale Malicious Code: A Research Agenda," DARPA, 2003.; Bruce Schneier, "Benevolent Worms," *Crypto-Gram*, 14 September 2003.; Frank Castaneda, Emre Can Sezer and Jun Xu. "WORM vs. WORM: Preliminary Study of an Active Counter-Attack Mechanism," *ACM workshop on Rapid Malcode (WORM), Washington, DC, 29 October 2004*.; Mason J. Molesky and Elizabeth A. Cameron, "Internet of Things: An Analysis and Proposal of White Worm Technology," *IEEE International Conference on Consumer Electronics (ICCE). Las Vegas, NV 11-13 January 2019*.

21 CERT organizations may exist under numerous bureaucratic frameworks that vary by state, and some are even operated by private sector actors. Here, however, we consider the functional role of independent national entities intended to coordinate response to ongoing cyber incidents for enterprise or sector level availability, integrity, and confidentiality objectives vice intended prosecution or intelligence objectives.

22 Clement Guitton, "Criminals and Cyber Attacks: The Missing Link between Attribution and Deterrence," *International Journal of Cyber Criminology* 6, no. 2 (July – December 2012): 1030–43.

23 Brian L. Owsley, "Beware of Government Agents Bearing Trojan Horses," *Akron Law Review* 48, no. 2 (2015).; Jonathan Mayer, "Government Hacking," *Yale Law Journal* 127, no. 3 (2017).; Eduardo R Mendoza. "Network Investigation Techniques: Government Hacking and the Need for Adjustment in the Third-Party Doctrine," *St Mary's Law Journal*, 49 (2017).; Christine W. Chen, "The Graymail Problem Anew in a World Going Dark: Balancing the Interests of the Government and Defendants in Prosecutions Using Network Investigative Techniques," *Columbia Science & Technology Review* XIX (Fall 2017).; Paul Ohm, "The Investigative Dynamics of the Use of Malware by Law Enforcement," *William & Mary Bill of Rights Journal* 26, no. 2 (2017).; Brian L. Owlsey, "Network Investigative Source Code and Due Process," *Digital Evidence and Electronic Signature Law Review* 14 (2017).

small part to continuing legal controversy, these actions are rarely highlighted in post-takedown case summaries.[24]

The dataset also omits routine takedown operations intended to counter ephemeral abuse and simple malicious hosting, as is commonly used in phishing, drive-by malware distribution, secondary payload staging, exfiltration drops, or other tactical functions by actors who anticipate prompt pressure upon use, and therefore are rotated with relatively high frequency. (The dynamics of this tactical level chase are well captured in the "Pyramid of Pain" analytic construct.)[25] Red-on-red cases are also likely underrepresented, due to limited observation and unwillingness of victims to provide any public disclosure.

Disruptive counter-cyber operations can be targeted across malware, command-and-control, and other supporting infrastructure, adversary operator freedom of action, or enabling transactional marketplaces. The differing nature of these defensive objectives plays a role in the effectiveness, or lack thereof, of disruptive counter-cyber disruptions. Simple prediction or even ready explanations of disruptive outcomes are clouded by these differing targeting objectives, sensitivity to initial conditions, and other case-specific factors.

TABLE 1: DATASET OF CYBER DISRUPTION EVENTS

#	Disruptive Event or Campaign	Approximate Date	Motivation of Disrupted Adversary	Disruption Actor
1	Anti-Christma Exec probable campaign[i]	December 1987	Vuln Reduction	Industry
2	Denzuko campaign targeting Brain[ii]	March 1988	Criminal	Red-On-Red
3	Cheese campaign targeting L1on[iii]	May 2001	Vuln Reduction	Unknown
4	CodeGreen campaign targeting CodeRed[iv]	September 2001	Vuln Reduction	Unknown
5	CRClean campaign targeting CodeRed[v]	September 2001	Vuln Reduction	Unknown
6	Klez campaign targeting CodeRed[vi]	October 2001	Criminal	Red-On-Red

24 Brian L. Owsley, "Beware of Government Agents Bearing Trojan Horses," *Akron Law Review* 48, no. 2. (2015).; Jonathan Mayer, "Government Hacking." *Yale Law Journal*, 127, no. 3. 2017.; Eduardo R Mendoza, "Network Investigation Techniques: Government Hacking and the Need for Adjustment in the Third-Party Doctrine," *St Mary's Law Journal*. 49 (2017).; Christine W. Chen, "The Graymail Problem Anew in a World Going Dark: Balancing the Interests of the Government and Defendants in Prosecutions Using Network Investigative Techniques," *Columbia Science & Technology Review* XIX (Fall 2017).; Paul Ohm, "The Investigative Dynamics of the Use of Malware by Law Enforcement," *William & Mary Bill of Rights Journal* 26, no. 2 (2017).; Brian L. Owlsey, "Network Investigative Source Code and Due Process," *Digital Evidence and Electronic Signature Law Review* 14 (2017).

25 David J. Bianco, "The Pyramid of Pain," 7 January 2014, http://detect-respond.blogspot.com/2013/03/the-pyramid-of-pain.html.

7	Columbia network worm vaccine architecture experiment[vii]	June 2003	Vuln Reduction	Industry
8	Welchia / Nachi campaign targeting Blaster[viii]	August 2003	Vuln Reduction	Unknown
9	Netsky campaign targeting Beagle and MyDoom[ix]	February 2004	Criminal	Red-On-Red
10	Shadowcrew – Carderplanet underground marketplace takedown[x]	November 2004	Criminal	Gov (LE)
11	Welchia / Nachi-B campaign targeting MyDoom[xi]	November 2004	Vuln Reduction	Unknown
12	Harbin "QBTP worm" experiment[xii]	August 2005	Vuln Reduction	Industry
13	eGold takedown[xiii]	December 2005	Criminal	Gov (LE)
14	RBN bulletproof hosting takedown[xiv]	November 2007	Criminal	Industry
15	Kraken botnet exploitation[xv]	April 2008	Criminal	Industry
16	Darkmarket underground marketplace takedown[xvi]	October 2008	Criminal	PPP
17	McColo bulletproof hosting takedown[xvii]	November 2008	Criminal	Industry
18	Changsha "P2P anti-worm" experiment[xviii]	November 2008	Vuln Reduction	Industry
19	Srizbi takedown attempt[xix]	November 2008	Criminal	Industry
20	Storm botnet exploitation[xx]	December 2008	Criminal	Industry
21	Conficker botnet disruption[xxi]	November 2008 to June 2010	Criminal	PPP
22	Torpig botnet exploitation[xxii]	January – February 2009	Criminal	Industry
23	Ghostnet exploitation & disclosure[xxiii]	March 2009	Espionage	Industry
24	Simulated Bluetooth proximity malware white worm experiment[xxiv]	April 2009	Vuln Reduction	Industry
25	3FN bulletproof hosting takedown[xxv]	June 2009	Criminal	Gov (LE)
26	Algiers "father worm" experiment[xxvi]	July 2009	Benevolent	Industry
27	"Independence Day" botnet exploitation[xxvii]	July 2009	Strategic Attack	Industry
28	Mega-D botnet takedown[xxviii]	November 2009	Criminal	Industry
29	Lethic takedown attempt[xxix]	January 2010	Criminal	Industry
30	Waledec (b49) takedown[xxx]	February 2010	Criminal	Industry
31	Mariposa takedown[xxxi]	February 2010	Criminal	PPP
32	Troyak bulletproof hosting takedown[xxxii]	March 2010	Criminal	Industry

33	Dumps.name / BadB underground marketplace disruption[xxxiii]	August 2010	Criminal	Gov (LE)
34	Pushdo / Cutwail botnet takedown[xxxiv]	August 2010	Criminal	Industry
35	Bredolab takedown[xxxv]	October 2010	Criminal	PPP
36	Rustock (b107) takedown[xxxvi]	March 2011	Criminal	Industry
37	Coreflood takedown[xxxvii]	April 2011	Criminal	LE
38	DNSChanger takedown[xxxviii]	November 2011	Criminal	PPP
39	Ice IX possible exploitation[xxxix]	June 2012	Criminal	Industry
40	Grum botnet takedown[xl]	July 2012	Criminal	PPP
41	UGNazi takedown[xli]	May 2012	Hacktivist & Criminal	Gov (LE)
42	Syrian Electronic Army DarkComet possible exploitation[xlii]	November 2012 onward	Espionage	Unknown
43	Brobot takedown[xliii]	January 2013	Strategic Attack	Unknown
44	Dexter POS malware possible exploitation[xliv]	February 2013	Criminal	Industry
45	APT1 disclosure[xlv]	February 2013	Espionage	Industry
46	APT1 exploitation[xlvi]	March 2013	Espionage	Industry
47	Kelihos takedown attempt[xlvii]	March 2013	Criminal	Industry
48	Liberty Reserve takedown[xlviii]	May 2013	Criminal	Gov (LE)
49	Citadel (b54) takedown[xlix]	June 2013	Criminal	Industry
50	Carberp exploitation[l]	June 2013	Criminal	Red-On-Red
51	Blackhole Exploit Kit sales disruption[li]	October 2013	Criminal	Gov (LE)
52	Silk Road underground marketplace takedown[lii]	October 2013	Criminal	Gov (LE)
53	Zeroaccess disruption[liii]	December 2013	Criminal	PPP
54	Blackshades takedown[liv]	May 2014	Criminal	Gov (LE)
55	APT2 / PUTTER PANDA disclosure[lv]	May 2014	Espionage	Industry
56	GameOverZeus takedown[lvi]	June 2014	Criminal & Espionage	PPP
57	Shylock / Hijack takedown[lvii]	July 2014	Criminal	Gov (LE)
58	Citadel possible exploitation[lviii]	August 2014	Criminal	Industry
59	“Operation Onymous” underground marketplace takedowns[lix]	November 2014	Criminal	Gov (LE)

60	Asprox disruption[lx]	January 2015	Criminal	Gov (LE)
61	Ramnit takedown attempt[lxi]	February 2015	Criminal	Gov (LE)
62	SIMDA takedown[lxii]	April 2015	Criminal	PPP
63	Neverquest / Vawtrack takedown[lxiii]	April 2015	Criminal	PPP
64	Beebone takedown[lxiv]	April 2015	Criminal	Gov (LE)
65	APT30* / NAIKON / OVERRIDE PANDA* / LOTUS PANDA* disclosure[lxv]	July 2015	Espionage	Industry
66	Opfake exploitation[lxvi]	September 2015	Criminal	Industry
67	Dridex takedown[lxvii]	October 2015	Criminal	Gov (LE)
68	Dirt Jumper / Drive / Pandora possible exploitation[lxviii]	October 2015	Criminal	Industry
69	Dyre disruption[lxix]	November 2015	Criminal	Gov (LE)
70	Dorkbot takedown[lxx]	December 2015	Criminal	PPP
71	Lurk / Angler disruption[lxxi]	June 2016	Criminal	Gov (LE)
72	Hajime campaign[lxxii]	October 2016	Vuln Reduction	Unknown
73	Avalanche / KOL takedown[lxxiii]	November 2016	Criminal	PPP
74	Nymaim disruption[lxxiv]	December 2016	Criminal	PPP
75	Chanitor distribution of Vawtrack disruption[lxxv]	January 2017	Criminal	Gov (LE)
76	Cerber / Sage exploitation[lxxvi]	February 2017	Criminal	Industry
77	Blackmoon exploitation[lxxvii]	March 2017	Criminal	Industry
78	Neutrino bot exploitation[lxxviii]	March 2017	Criminal	Industry
79	Gaudox bot exploitation[lxxix]	March 2017	Criminal	Industry
80	Kelihos takedown[lxxx]	April 2017	Criminal	PPP
81	Brickerbot campaign[lxxxi]	April 2017	Vuln Reduction	Unknown
82	APT3 / GOTHIC PANDA / UPS TEAM disclosure[lxxxii]	May 2017	Espionage	Unknown
83	Plug-X possible exploitation[lxxxiii]	June 2017 onward	Espionage	Unknown
84	AlphaBay and Hansa underground marketplace takedowns[lxxxiv]	July 2017	Criminal	Gov (LE)
85	WireX disruption[lxxxv]	August 2017	Criminal	Industry
86	Andromeda botnet takedown[lxxxvi]	November 2017	Criminal	Gov (LE)

87	Mirai botnet disruption[lxxxvii]	March 2018	Hacktivist & Criminal	Gov (LE)
88	MaxiDed bulletproof hosting takedown[lxxxviii]	May 2018	Criminal	Gov (LE)
89	VPNFilter takedown[lxxxix]	May 2018	Espionage & Strategic Attack	PPP
90	MegaladonHTTP botnet possible exploitation[xc]	June 2018	Criminal	Industry
91	APT10 / STONE PANDA / MenuPass / POTASSIUM disclosure[xci]	August 2018	Espionage	Unknown
92	3ve takedown[xcii]	October 2018	Criminal	Gov (LE)
93	VPNFilter possible exploitation[xciii]	November 2019	Espionage & Strategic Attack	Unknown
94	Joanap takedown[xciv]	January 2019	Espionage & Strategic Attack	Gov (LE)
95	COBALT STRIKE abuse disclosure[xcv]	February 2019	Espionage & Criminal	Industry
96	Abdallah / Yalishanda hosting takedown[xcvi]	July 2019	Criminal	Gov (LE)
97	Retadup takedown[xcvii]	August 2019	Criminal	PPP
98	APT34 / HELIX KITTEN / OILRIG / COBALT GYPSY / CHRYSENE disclosure[xcviii]	April – May 2019	Espionage & Strategic Attack	Unknown
99	APT17 / AURORA PANDA / DOGFISH disclosure[xcix]	July 2019	Espionage	Unknown
100	CyberBunker bulletproof hosting takedown[c]	September 2019	Criminal	Gov (LE)
101	Turla / VENOMOUS BEAR / KRYPTON compromise of APT34 / OILRIG / CHRYSENE[ci]	November 2019	Espionage	Red-On-Red
102	H-Worm possible exploitation[cii]	November 2019	Espionage	Unknown
103	APT33 / REFINED KITTEN / COLBAT TRINITY infrastructure disclosure[ciii]	November 2019	Espionage	Industry

5. INITIAL INSIGHTS AND CONCLUSION

The debate over the appropriate approach, timing, and manner of actions intended to deny and degrade ongoing cyber threats closer to their origins has to date been a largely theoretical affair. The disconnects between policy communities and the operators and researchers engaged in the day-to-day fight on the wire have meant that in many cases, well-intentioned thinkers on both sides have been effectively talking past each other when discussing concepts of operation, desired end states,

and perceived drawbacks. While many key details of current and proposed future operations remain locked in classified discourse, the development of the framework proposed here, and the underlying dataset which has informed it, demonstrate that there is indeed a robust record of prior incidents by which to nominate courses of actions, illuminate conflicting equities, and advance reasoned arguments for both sides. Grounding ongoing conversations using a publicly documented dataset and the associated analytical features of these identified case studies will be useful in improving debates over differing policy and technical proposals.

Initial cross-case analysis already offers preliminary insights and clarifies questions to be further explored in depth for more robust testing and validation. Commonalities across the entirety of the case dataset importantly suggest that operational disruption is rarely accomplished as a single decisive action, at least where adversary operators, developers, and planners continue to enjoy a sustained base of uninterrupted support. However, merely because a single action will not render the adversary *hors de combat* does not negate the utility of disruption. Forcing adversary adaptation may add value, particularly where such a response requires investment disproportionate to the value of continuing operations or where adversary resourcing may be constrained in some other dimensions. Here the bias of the extant cases in the dataset must also be considered, where more technically effective options to achieve decisive results against adversary operations may have been available but precluded by the decision to pursue the operation under a law-enforcement framework, as opposed to national security or military authorities. From these cases, it appears that simpler direct-action options may have been available to disrupt adversary targets, but that more complex (as well as likely therefore more fragile) and higher-risk operations were conducted in order to preserve evidence for prosecution purposes, or to serve civil and coordination processes for later remediation of compromised victim systems. One may not presume that all future disruption efforts will be so constrained.

These potential issues have substantial relevance where targets are transnational criminal networks, especially those involved in both criminal activity and espionage operations, as a proxy on behalf of hostile intelligence services. In operational practice, one might note the case of the August 2019 Retadup takedown in light of these issues. During this law-enforcement action, pursued under European jurisdiction, disruption operators took steps to remove malware from compromised victim systems after successful takeover of botnet command & control infrastructure – measures that had been precluded in a number of previous campaigns conducted under other jurisdictions for fear of liability exposure, or due to ethical concerns.[26]

26 Felix Leder, Tillmann Werner, and Peter Martini, "Proactive Botnet Countermeasures – An Offensive Approach," in *The Virtual Battlefield: Perspectives on Cyber Warfare*, ed. Christian Czosseck and Kenneth Geers. IOS Press, 2009.; David Dittrich, Felix Leder, and Tillmann Werner, "A Case Study in Ethical Decision Making Regarding Remote Mitigation of Botnets," *International Conference on Financial Cryptography and Data Security. Tenerife, Canary Islands, Spain, 25-28 January 2010*.; Sam Zeitlin, "Botnet Takedowns and the Fourth Amendment," *New York University Law Review* 90, no. 2 (May 2015): 746-778.

Disruptions made for vulnerability reduction, such as the CodeRed / CodeGreen case study, are important transitional actions, often taken by defense-minded actors who have expressed frustration at unmitigated exposure or ongoing adversary action that has apparently gone unaddressed – and including what are believed to be the earliest documented hackback actions by non-state actors. These cases serve as an instructive contrast to more structured and deliberate operations as they are typically unilateral, with uncoordinated execution and significant potential for collateral damage and escalation. Experimentation with purely technical capabilities, they may have informed later concepts of operation by other actors acting within different constraints, where technical options are modified to meet acceptable criteria defined by political, judicial, or operational oversight.

Red-on-red cases surface with particular salience where state intelligence services – in an attempt to advance deception themes or achieve surprise – leverage criminal capabilities acquired through transactional engagements, or coercive leverage, to intertwine espionage and strategic attack objectives with criminal operations. Where such capabilities are co-mingled, the state service involved takes on a greater operational risk, as criminal infrastructure is more commonly targeted by other criminals who share common understanding of tactics, techniques, and procedures and are aware of routine failures in operational practice that may lead to takeover or competitive disruption. Yet at the same time, these cases perhaps suggest that escalation concerns over adversary reaction to disruptive operations may be lessened in a number of situations, given prior incidents in which state actors leveraging co-mingled infrastructure apparently did not respond directly. Nonetheless, the small number of documented incidents demands further cautious consideration beyond such tentative, preliminary insight.

We hope this framework and dataset bring transparency and repeatability to the critical issue of disruptive counter-cyber operations. Future research in this area – both by academics and practitioners in the government or commercial cyber threat intelligence field – should improve our framework, apply it to the full data set to allow deeper insights, and develop additional case studies. A student capstone project at Columbia University's School of International and Public Affairs is specifically researching the impact on adversaries of one specific kind of disruption, public disclosure.

The dataset published here does demonstrate that, far from an unprecedented break with past practice, new proposed disruption approaches may be considered evolutionary in design and execution and may be evaluated within a common framework. Lessons from prior disruptive actions can improve future operations by the US government, its allies, and other likeminded actors – especially given reported intentions to pursue more assertive employment of offensive measures for counter-cyber operations within

the context of the persistent engagement framework, implementing the strategic vision of "defend forward". These cases help to understand the likely upper bounds of such operations, and how such actions may be tailored to cause the most friction under differing situations. It is believed that a neutral, objective analytic construct is the best mechanism for evaluating comparative countering options, with the hope that it will focus planning and action in a manner that denies and degrades adversary capabilities at the greatest scale, the least cost, and with the lowest chance of escalation.

APPENDIX: REFERENCES FOR TABLE 1, DATASET OF CYBER DISRUPTION EVENTS

i Capek, Peter G., David M. Chess, Steve R. White, and Alan Fedeli. "Merry Christma: An Early Network Worm." *IEEE Security & Privacy* 1, no. 5 (Sept.–Oct. 2003): 26-34.

ii Skulason, Fridrik. "The Search for Den Zuk." *Virus Bulletin*. February 1991.

iii Barber, Bryan. "Cheese Worm: Pros and Cons of a Friendly Worm." SANS Institute. 26 July 2001.

iv Der HexXer, Herbert. "CodeGreen Beta Release." *Vuln-Dev*. 1 September 2001.

v Metzer, David J. "The Coming Age of Defensive Worms." *Toorcon*. September 2003.

vi Symantec. "W32.Klez.A@mm." 25 October 2001.

vii Sidiroglou, Stelios, Angelos D. Keromytis. "A Network Worm Vaccine Architecture." *Proceedings of the 12th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE). Linz, Austria. 11-11 June 2003*.

viii Symantec. "W32.Welchia.Worm." 18 August 2003.

ix Edwards, Dwayne. "Netsky.p Mass Mailer Worm Analysis." *SANS Institute*. 9 January 2005.

x DOJ. "Nineteen Individuals Indicted in Internet 'Carding' Conspiracy: Shadowcrew Organization Called 'One-Stop Online Marketplace for Identity Theft."" 28 October 2004. https://www.justice.gov/archive/criminal/cybercrime/press-releases/2004/mantovaniIndict.htm; James Verini. "The Great Cyberheist." *The New York Times Magazine*. 10 November 2010.; DOJ. "Ukrainian National Who Co-Founded Cybercrime Marketplace Sentenced To 18 Years in Prison." 12 December 2013, https://www.justice.gov/usao-edny/pr/ukrainian-national-who-co-founded-cybercrime-marketplace-sentenced-18-years-prison.

xi Symantec. "W32.Welchia.B.Worm." 18 November 2004.

xii Liu, Yi-Xuan, Xiao-Chun Yun, Bai-Ling Wang, Hai-Bin Sun. "QBTP Worm: An Anti-Worm with Balanced Tree Based Spreading Strategy." *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics. Guangzhou, People's Republic of China. 18-21 August 2005*.

xiii Zetter, Kim. "Bullion and Bandits: The Improbable Rise and Fall of E-Gold." Wired. 9 June 2009.

xiv iDefense. "The Russian Business Network: Rise and Fall of a Criminal ISP." 3 March 2008.

xv Mushtaq, Atif. "Kraken Botnet – A Detailed Analysis." *FireEye*. 17 April 2008, https://www.fireeye.com/blog/threat-research/2008/04/kraken-botnet-1.html.

xvi Alperovitch, Dmitri and Keith Mularski. "Fighting Russian Cybercrime Mobsters: Report from the Trenches." *Black Hat USA. Las Vegas, Nevada. 29-30 July 2009*.

xvii Krebs, Brian. "Host of Internet Spam Groups Is Cut Off." *Washington Post*. 12 November 2008.

xviii Wang, Bin, Piao Ding, Jinfang Sheng. "P2P Anti-worm: Modeling and Analysis of a New Worm Counter-measurement Strategy." *9th International Conference for Young Computer Scientists. Hunan, People's Republic of China. 18-21 November 2008*.

xix Keizer, Gregg. "Massive Botnet Returns From The Dead, Starts Spamming." Computerworld. 26 November 2008.

xx Wicherski, Georg 'oxff', Tillmann Werner, Felix Leder, Mark Schlösser. "Stormfucker: Owning the Storm Botnet." *25th Chaos Communication Congress. Berlin, Germany. 29 December 2008*.

xxi Rendon Group. "Conficker Working Group: Lessons Learned." January 2011.

xxii Stone-Gross, Brett, Marco Cova, Lorenzo Cavallaro, Bob Gilbert, Martin Szydlowski, Richard Kemmerer, Christopher Kruegel, Giovanni Vigna. "Your Botnet is my Botnet: Analysis of a Botnet Takeover." *16th ACM conference on Computer and communications security (CCS). Chicago, Illinois. 9-13 October 2009*.

xxiii Information Warfare Monitor. "Tracking GhostNet: Investigating a Cyber Espionage Network." 29 March 2009.

xxiv Zyba, Gjergji, Geoffrey M. Voelker, Michael Liljenstam, Andras Mehes, Per Johansson. "Defending Mobile Phones from Proximity Malware." *IEEE INFOCOM. Rio de Janeiro, Brazil. 19-25 April 2009*.

xxv Krebs, Brian. "FTC Sues, Shuts Down N. Calif. Web Hosting Firm." *Washington Post*. 4 June 2009.

xxvi Berbar, Ahmed, Mohamed Ahmednacer. "Testing and Fault Tolerance Approach for Distributed Software Systems Using Nematode Worms." *Proceedings of the 4th International Conference on Queueing Theory and Network Applications (QTNA). July 2009*.

xxvii iSIGHT Partners. "Peer-to-Peer Command-and-Control Architecture Likely Used in Sustained DDoS Attacks Against South Korean and U.S. Targets." 9 July 2009.; Nguyen, Minh Duc. "Comments on Korea and US DDOS Attacks." *BKIS*. 14 July 2009, http://blog.bkis.com/?p=718.

xxviii Lin, Phillip. "Anatomy of the Mega-D takedown." Network Security. (December 2009): 4-7.; Cho, Chia Yuan, Juan Caballero, Chris Grier, Vern Paxson, and Dawn Song. "Insights from the Inside: A View of Botnet Management from Infiltration." *3rd Usenix Workshop on Large Scale Exploits and Emergent Threats (LEET). San Jose, CA. 28-30 April 2010*.

xxix Zscaler. "Lethic Botnet Returns, Uses "Realtek" Identifier." 10 November 2010.

xxx Cranton, Tim. "Cracking Down on Botnets." *Microsoft*. 24 February 2010, https://blogs.technet.microsoft.com/microsoft_on_the_issues/2010/02/24/cracking-down-on-botnets/.

xxxi Sully, Matt, and Matt Thompson. "The Deconstruction of the Mariposa Botnet." *Defence Intelligence.* February 2010.

xxxii McMillan, Robert. "Zeus Botnet Dealt a Blow as ISP Troyak Knocked Out." *Computerworld*. 10 March 2010; McMillan, Robert. "After Takedown, Botnet-Linked ISP Troyak Resurfaces." *Computerworld*. 10 March 2010.

xxxiii DOJ. "Alleged International Credit Card Trafficker Arrested in France on U.S. Charges Related to Sale of Stolen Card Data." 11 August 2010, https://www.justice.gov/opa/pr/alleged-international-credit-card-trafficker-arrested-france-us-charges-related-sale-stolen.

xxxiv iSIGHT Partners. "Potential 'Dead Hand' C&C Architecture Suggested by Adversary Adaptation Following Failed Botnet Takedown Attempt." 11 February 2010.; JD Work. "Autonomy & Conflict Management in Offensive & Defensive Cyber Engagement." *IWCon. Nashville, TN. 5-7 April 2016*.

xxxv Williams, Jeff. "Bredolab Takedown, Another Win for Collaboration." *Microsoft*. 26 October 2010. http://blogs.technet.com/b/mmpc/archive/2010/10/26/bredolabtakedown-another-win-for-collaboration.aspx.

xxxvi Boscovich, Richard. "Taking Down Botnets: Microsoft and the Rustock Botnet." *Microsoft*. 17 March 2011, https://blogs.technet.microsoft.com/microsoft_on_the_issues/2011/03/17/taking-down-botnets-microsoft-and-the-rustock-botnet/.

xxxvii DOJ. "Department of Justice Takes Action to Disable International Botnet." 13 April 2011, https://www.justice.gov/opa/pr/department-justice-takes-action-disable-international-botnet.

xxxviii Trend Micro. "OPERATION GHOST CLICK: The Rove Digital Takedown." 2012.

xxxix Sood, Aditya K. "For Fun - XSS in ICE IX C&C Panel." 12 June 2012, https://secniche.blogspot.com/2012/06/for-fun-xss-in-ice-ix-bot-admin-panel.html.

xl Mushtaq, Atif. "Grum, World's Third-Largest Botnet, Knocked Down." *FireEye*. 18 July 2012.

xli Honan, Mat. "Cosmo, the Hacker 'God' Who Fell to Earth." *Wired*. 11 September 2012.

xlii Denbow, Shawn and Jesse Hertz. "Pest Control: Taming the Rats." *Matasano Security*. October 2012.

xliii Work, JD. "Echoes of Ababil: Re-Examining Formative History of Cyber Conflict and its Implications for Future Engagements." *Society of Military History Annual Conference. Cincinnati, OH. 9-12 May 2019*.

xliv Wallace, Brian. "A Study in Bots: Dexter." *Cylance*. 14 March 2014, https://threatvector.cylance.com/en_us/home/a-study-in-bots-dexter-pos-botnet-malware.html.

xlv Mandiant. "APT1: Exposing One of China's Cyber Espionage Units." 19 February 2013, https://www.fireeye.com/content/dam/fireeye-www/services/pdfs/mandiant-apt1-report.pdf.

xlvi Rascagnères, Paul. "APT1: Technical Backstage." *Malware.lu*. 27 March 2013.

xlvii Werner, Tillmann. "Peer-to-Peer Poisoning Attack against the Kelihos.C Botnet." *CrowdStrike*. 21 March 2013.

xlviii DOJ. "Manhattan U.S. Attorney Announces Charges Against Liberty Reserve, One of World's Largest Digital Currency Companies, and Seven of its Principals and Employees for Allegedly Running a $6 Billion Money Laundering Scheme." 28 May 2013, https://www.justice.gov/usao-sdny/pr/manhattan-us-attorney-announces-charges-against-liberty-reserve-one-world-s-largest.

xlix Meisner, Jeffrey. "Microsoft Works with Financial Services Industry Leaders, Law Enforcement and Others to Disrupt Massive Financial Cybercrime Ring." *Microsoft*. 5 June 2013.

l Steven K. “Carberp Remote Code Execution: Carpwned.” XyliBox blog. 28 June 2013, https://www.xylibox.com/2013/06/carberp-remote-code-execution-carpwned.html.

li FireEye. “Black Hole Exploit Kit: The Rise and Fall of an Exploit Kit Giant.” 28 March 2014.

lii Zetter, Kim. “How the Feds Took Down the Silk Road Drug Wonderland.” *Wired*. 18 November 2013, https://www.wired.com/2013/11/silk-road/.

liii Microsoft. “Microsoft, the FBI, Europol and Industry Partners Disrupt the Notorious ZeroAccess Botnet.” 5 December 2013, https://news.microsoft.com/2013/12/05/microsoft-the-fbi-europol-and-industry-partners-disrupt-the-notorious-zeroaccess-botnet/.

liv FBI. “International Blackshades Malware Takedown Coordinated Law Enforcement Actions Announced.” 19 May 2014, https://www.fbi.gov/news/stories/international-blackshades-malware-takedown-1.

lv CrowdStrike. “Putter Panda.” May 2014.

lvi DOJ. “U.S. Leads Multi-National Action Against ‘Gameover Zeus’ Botnet and ‘Cryptolocker’ Ransomware, Charges Botnet Administrator.” 2 June 2014, https://www.justice.gov/opa/pr/us-leads-multi-national-action-against-gameover-zeus-botnet-and-cryptolocker-ransomware.

lvii Europol. “Global Action Targeting Shylock Malware.” 10 July 2014, https://www.europol.europa.eu/newsroom/news/global-action-targeting-shylock-malware.

lviii Sood, Aditya K. “Exploiting Fundamental Weaknesses in Botnet Command and Control Panels.” *Black Hat. Las Vegas, NV. 2-7 August 2014*.

lix Europol. “Global Action Against Dark Markets on Tor Network.” 7 November 2014, https://ec.europa.eu/home-affairs/what-is-new/news/news/2014/20141107_01_en.

lx Secureworks. “Evolution of the GOLD EVERGREEN Threat Group.” 15 May 2017.

lxi Europol. “Botnet Taken Down Through International Law Enforcement Cooperation”. 25 February 2015. https://www.europol.europa.eu/newsroom/news/botnet-taken-down-through-international-law-enforcement-cooperation; Symantec. “Ramnit Cybercrime Group Hit by Major Law Enforcement Operation.” 25 February 2015.; Trend Micro. “Ramnit: The Comeback Story of 2016.” 20 February 2017.

lxii Interpol. “INTERPOL Coordinates Global Operation to Take Down Simda Botnet.” 13 April 2015, https://www.interpol.int/en/News-and-Events/News/2015/INTERPOL-coordinates-global-operation-to-take-down-Simda-botnet.

lxiii Takada, Kazuki. “Behind Operation Banking Malware Takedown and the Progression of Malware Sophistication.” *Code Blue. Tokyo, Japan. 20-21 October 2016*.

lxiv Europol. “International Police Operation Targets Polymorphic Beebone Botnet.” 9 April 2015, https://www.europol.europa.eu/newsroom/news/international-police-operation-targets-polymorphic-beebone-botnet.

lxv FireEye. “APT 30 and the Mechanics of a Long-Running Cyber Espionage Operation.” 12 April 2015.; ThreatConnect. “Project CameraShy: Closing the Aperture on China’s Unit 78020.” July 2015.

lxvi Huang, Wayne and Sun Huang. “24 Techniques to Gather Threat Intel and Track Actors.” *Black Hat Asia. Singapore. 28-31 March 2017*.

lxvii Secureworks. “Dridex (Bugat v5) Botnet Takeover Operation.” 13 October 2015.

lxviii Watkins, Lanier, Kurt Silberberg, Jose Andre Morales, William H. Robinson. “Using Inherent Command and Control Vulnerabilities to Halt DDoS Attacks.” *10th International Conference on Malicious and Unwanted Software. Fajardo, Puerto Rico. 20–22 October 2015*.

lxix Symantec. “Dyre: Operations of Bank Fraud Group Grind to Halt Following Takedown.” 8 February 2016.

lxx Interpol. “INTERPOL Supports Global Operation Against Dorkbot Botnet.” 4 December 2015.

lxxi Biasini, Nick. “Connecting the Dots Reveals Crimeware Shake-up.” *Cisco TALOS*. 7 July 2016.

lxxii Grange, Waylon. “Hajime Worm Battles Mirai for Control of the Internet of Things.” *Symantec*. 18 April 2017.; Yamaguchi, Shingo, Pattara Leelaprute. “Hajime Worm with Lifespan and Its Mitigation Evaluation Against Mirai Malware Based on Agent-Oriented Petri Net PN2.” *IEEE International Conference on Consumer Electronics (ICCE). Las Vegas, NV 11–13 January 2019*.

lxxiii Europol. “Avalanche Network Dismantled in International Cyber Operation.” 1 December 2016, https://www.europol.europa.eu/newsroom/news/%E2%80%98avalanche%E2%80%99-network-dismantled-in-international-cyber-operation.

lxxiv CrowdStrike. “Goznym: Gozi Malware Hybrid Bundled with the Nymaim Loader.” 2 June 2017.

lxxv FireEye. “Operational Net Assessment for Cyber Crime: January to March 2017.” 4 April 2017.

lxxvi Huang and Huang, 2017.

lxxvii Huang and Huang, 2017.

lxxviii Huang and Huang, 2017.

lxxix Huang and Huang, 2017.

lxxx CrowdStrike. “Inside the Takedown of ZOMBIE SPIDER and the Kelihos Botnet.” 13 April 2017.

lxxxi DHS CISA. “BrickerBot Permanent Denial-of-Service Attack.” 12 April 2017, https://www.us-cert.gov/ics/alerts/ICS-ALERT-17-102-01A.

lxxxii Recorded Future. “Recorded Future Research Concludes Chinese Ministry of State Security Behind APT3.” 17 May 2017.; Checkpoint. “UPSynergy: Chinese-American Spy vs. Spy Story.” 5 September 2019.

lxxxiii Grange, Waylon. “Digital Vengeance: Exploiting the Most Notorious C&C Toolkits.” *Black Hat. Las Vegas, NV. 22–27 July 2017*.

lxxxiv Europol. “Massive Blow to Criminal Dark Web Activities After Globally Coordinated Operation.” 20 July 2017, https://www.europol.europa.eu/newsroom/news/massive-blow-to-criminal-dark-web-activities-after-globally-coordinated-operation.

lxxxv Grooten, Martijn. “WireX DDoS Botnet Takedown Shows the Best Side of the Security Industry.” *Virus Bulletin*. 29 August 2017.

lxxxvi Europol. “Andromeda Botnet Dismantled in International Cyber Operation.” 4 December 2017, http://www.eurojust.europa.eu/press/PressReleases/Pages/2017/2017-12-04.aspx.

lxxxvii Graff, Garrett M. “The Mirai Botnet Architects Are Now Fighting Crime With the FBI.” *Wired*. 18 September 2018.

lxxxviii Noroozian, Arman, Eelco van Veldhuizen, Carlos H. Ganan, Sumayah Alrwais, Damon McCoy, Michel van Eeten. “Platforms in Everything: Analyzing Ground-Truth Data on the Anatomy and Economics of Bullet-Proof Hosting.” *28th Usenix Security Symposium. Santa Clara, CA. 14–16 August 2019*.

lxxxix DOJ. “Justice Department Announces Actions to Disrupt Advanced Persistent Threat 28 Botnet of Infected Routers and Network Storage Devices.” 23 May 2018, https://www.justice.gov/opa/pr/justice-department-announces-actions-disrupt-advanced-persistent-threat-28-botnet-infected.

xc Nachum, Shay, Assaf Schuster, Opher Etzion. “Detection in the Dark – Exploiting XSS Vulnerability in C&C Panels to Detect Malwares.” *Cyber Security Cryptography and Machine Learning (CSCML). Beer Sheva, Israel. 21-22 June 2018*.

xci FireEye. “Assessment of Recent Public Reports Regarding APT10.” 8 August 2019.

xcii DHS CISA. “3ve – Major Online Ad Fraud Operation.” 27 November 2018, https://www.us-cert.gov/ncas/alerts/TA18-331A.

xciii Olney, Matthew. “Seeing Broad Scanning...” *Twitter*. 27 November 2018, https://twitter.com/kpyke/status/1068141372543242240.

xciv DOJ. “Justice Department Announces Court-Authorized Efforts to Map and Disrupt Botnet Used by North Korean Hackers.” 30 January 2019, https://www.justice.gov/opa/pr/justice-department-announces-court-authorized-efforts-map-and-disrupt-botnet-used-north.

xcv Strategic Cyber, LLC. “Cobalt Strike Team Server Population Study.” 19 February 2019. https://blog.cobaltstrike.com/2019/02/19/cobalt-strike-team-server-population-study/; Work, JD. “In Wolf’s Clothing: Complications of Threat Emulation in Contemporary Cyber Intelligence Practice.” *Cyber Incident. University of Oxford, 3-4 June 2019*.

xcvi Krebs, Brian. “Meet the World’s Biggest ‘Bulletproof’ Hoster.” *Krebs on Security*. 16 July 2019. https://krebsonsecurity.com/2019/07/meet-the-worlds-biggest-bulletproof-hoster/; Security Service of Ukraine. “SBU Jointly with Foreign Colleagues Blocks Activity of Powerful Hacker Group.” 16 July 2019, https://ssu.gov.ua/en/news/1/category/21/view/6281#.J1jZcicu.dpbs.

xcvii Vojtěšek, Jan. “Putting an End to Retadup: A Malicious Worm that Infected Hundreds of Thousands.” *Avast*. 28 August 2019, https://decoded.avast.io/janvojtesek/putting-an-end-to-retadup-a-malicious-worm-that-infected-hundreds-of-thousands/.

xcviii FireEye. “Leaking Campaigns Designed to Degrade Iranian Cyber Capabilities Continue.” 11 June 2019.

xcix FireEye. “APT17 Outed as MSS Operation.” 25 July 2019.

c Krebs, Brian. “German Cops Raid “Cyberbunker 2.0,” Arrest 7 in Child Porn, Dark Web Market Sting.” *Krebs on Security*. 28 September 2019, https://krebsonsecurity.com/2019/09/german-cops-raid-cyberbunker-2-0-arrest-7-in-child-porn-dark-web-market-sting/.

ci National Security Agency and National Cyber Security Center. “Turla Group Exploits Iranian APT to Expand Coverage of Victims.” 21 October 2019, https://www.ncsc.gov.uk/news/turla-group-exploits-iran-apt-to-expand-coverage-of-victims.

cii Reporting by the self-styled “Democratic People’s Republic of Korea Computer Emergency Response Team (CERT).” “Vulnerability in the Remote Administration Tool (RAT) H-Worm.” 6 November 2019.

ciii Hacquebord, Feike, Cedric Pernet, and Kenney Lu. “More than a Dozen Obfuscated APT33 Botnets Used for Extreme Narrow Targeting.” *Trend Micro*. 13 November 2019, https://blog.trendmicro.com/trendlabs-security-intelligence/more-than-a-dozen-obfuscated-apt33-botnets-used-for-extreme-narrow-targeting/.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Using Global Honeypot Networks to Detect Targeted ICS Attacks

Michael Dodson
PhD Candidate
Department of Computer Science and Technology
University of Cambridge
Cambridge, United Kingdom
md403@cam.ac.uk

Alastair R. Beresford
Professor of Computer Security
Department of Computer Science and Technology
University of Cambridge
Cambridge, United Kingdom
arb33@cam.ac.uk

Mikael Vingaard
Industrial Security Researcher
Industrial Defenica
SecuriOT
Copenhagen, Denmark
info@honeypot.dk

Abstract: Defending industrial control systems (ICS) in the cyber domain is both helped and hindered by bespoke systems integrating heterogeneous devices for unique purposes. Because of this fragmentation, observed attacks against ICS have been targeted and skilled, making them difficult to identify prior to initiation. Furthermore, organisations may be hesitant to share business-sensitive details of an intrusion that would otherwise assist the security community.

In this work, we present the largest study of high-interaction ICS honeypots to date and demonstrate that a network of internet-connected honeypots can be used to identify and profile targeted ICS attacks. Our study relies on a network of 120 high-interaction honeypots in 22 countries that mimic programmable logic controllers and remote terminal units. We provide a detailed analysis of 80,000 interactions over 13 months, of which only nine made malicious use of an industrial protocol. Malicious interactions included denial of service and replay attacks that manipulated

logic, leveraged protocol implementation gaps and exploited buffer overflows. While the yield was small, the impact was high, as these were skilled, targeted exploits previously unknown to the ICS community.

By comparison with other ICS honeypot studies, we demonstrate that high-quality deception over long periods is necessary for such a honeypot network to be effective. As part of this argument, we discuss the accidental and intentional reasons why an internet-connected honeypot might be targeted. We also provide recommendations for effective, strategic use of such networks.

Keywords: *honeypot, industrial control system, ICS*

1. INTRODUCTION

Industrial Control Systems (ICS) are used to command, manage, or regulate devices or physical systems in industry (e.g., chemical processing), infrastructure (e.g., power generation), and building automation (e.g., fire suppression). Devices communicate using ICS-specific protocols, most of which are legacy point-to-point or broadcast protocols designed with the assumption that devices are connected with dedicated cabling; however, many of these protocols are now layered on top of ethernet and TCP or UDP, and devices use existing IP-based networks, including the internet, to communicate.

ICS security has not kept up with this growing digitisation and connectivity. The proprietary nature of most industrial software and the relatively low profile of industrial devices result in limited vulnerability hunting and disclosure [1] – [3]. For example, all versions of the two most popular proprietary (VxWorks) and open-source (FreeRTOS) real-time operating systems (RTOSes) have a total of 54 entries in the National Vulnerability Database (NVD) at the time of writing, compared with over 2,000 records for Windows 10 and over 800 records for Ubuntu 18.04. Further, all 'critical' VxWorks vulnerabilities in the NVD came from a single disclosure. Similarly, all but two of the FreeRTOS vulnerabilities came from a single disclosure. In each case, security researchers found more than 10 vulnerabilities that allowed remote code execution, data leakage, and denial of service attacks. Most were memory safety vulnerabilities and had existed in the software for more than a decade. Because these RTOSes are highly configurable, it is hard to estimate the number of affected devices; however, it is likely to exceed two billion [1], [4]. For comparison, the initial install target for Windows 10 was only one billion devices [5]. Further, when vulnerabilities

are identified, the industrial community demonstrates a strong resistance to patching, partly due to the high cost of regression testing and recertification by both the vendor and user [6]. Additionally, industrial networks have limited host-based security or logging opportunities, complicating forensic efforts. Even when forensic examination is possible, industrial network compromises are generally business-sensitive, so post-exploit forensic efforts rarely result in public disclosure of vulnerabilities, though ICS security companies often publish summary reports, such as those for Triton/Trisis [7]. Finally, few industrial protocols employ authentication or encryption; therefore, ICS devices will consider any well-formed packet to be valid, including those that request information or command changes of state [8], allowing malicious manipulation of device behaviour without actually exploiting any specific vulnerability. Together, these factors result in a vulnerable industrial environment and create unique security challenges.

Successful attacks against ICS have all targeted specific organisations and devices (e.g., Stuxnet [9], Triton/Trisis [7], CRASHOVERRIDE [10]) or have targeted vendors directly (e.g., [11]); therefore, unlike other domains where attacks are large-scale and indiscriminate, such as the Internet of Things (IoT) domain, there are limited means for researchers to gather open-source intelligence on ICS attack methods, motivations and campaigns. In domains such as IoT, honeypots have been effective tools to track and profile malicious behaviour [12], but they rely on either indiscriminate or easily deceived attackers, neither of which apply to current ICS adversaries. To date, the use of ICS honeypots for security research has been largely limited to monitoring internet-wide scanning.

Despite these challenges, we show that a geographically distributed network of high-interaction ICS honeypots can be an effective tool for identifying and profiling new, targeted attacks against ICS devices. We make the following contributions in this paper:

- A description of the largest, high-interaction ICS honeypot study to date.
- A discussion of multiple, new ICS exploits (zero days) identified by the honeypot network.
- An assessment of the growing overlap between ICS and IoT-aware scanning and botnet infections.
- An explanation of the limitations of previous ICS honeypot studies and recommendations for successful networks of ICS honeypots for security research.

2. BACKGROUND

A. Honeypots

Honeypots are computer security systems that emulate production systems and either decoy attackers away from the production system, provide warning of an intrusion, or allow attacker behaviour to be studied [12], [13]. Honeypots have been designed to emulate individual computers, such as laptops, servers, IoT and ICS devices [12], [14], and larger systems, such as electrical substations [15]. As a security device, they can be used as part of a defence-in-depth strategy alongside anti-virus software, segmented networks and firewalls. As a research tool, they are often used as stand-alone devices directly connected to the internet.

Honeypots can be characterised by their purpose and level of interaction [12]. The *purpose of interaction* refers to whether the honeypot is part of a production system, designed as part of a security solution for a given network or device, or a research device designed to attract attackers and study their behaviour [16]. The *level of interaction* refers to how well the honeypot emulates the target device, which determines how easy it is for the attacker to identify that they are interacting with a honeypot. The level of interaction is generally categorised as low, medium, or high, though these categories are not well-defined. A low-interaction honeypot may be a simple script that only emulates a login screen but no stateful device behaviour. A high-interaction honeypot may be an actual device or system, not an emulation, which is instrumented to record details of attacker behaviour on the system [17], [18].

Because honeypots have no purpose on a network except to deceive potential attackers, any interaction by an attacker with such a honeypot demonstrates that the attacker either lacks knowledge or is indiscriminate. If an attacker has sufficient knowledge and a specific target, then they can interact directly with the target device on a network and leave any honeypots untouched. If the attacker has less knowledge, but still has a specific target, they may have to scan a network to find the target device. In this case, they will interact with the honeypot and notify the defender of the attacker's presence, even if the attacker is able to avoid further interaction with the honeypot. In a less discriminate scenario, where the attacker is looking for any vulnerable device, they may go further and continue to interact with the honeypot, attempting to exploit vulnerabilities. Therefore, internet-connected, research honeypots have been effectively used to detect and monitor large-scale, indiscriminate attacks [12], but not knowledgeable, targeted attacks [19], [20].

Within the ICS community, there are several, open-source honeypots available. Conpot is a low-interaction honeypot capable of responding accurately to network scans [14]. It is easy to set up and scales well, making it a good candidate to research internet-

wide scanning [19] – [21]; however, its inability to interact with an attacker limits its utility in detecting and characterising ICS attacks, and studies using Conpot have yet to identify any new or targeted ICS attacks [19] – [21]. MiniCPS is a framework for higher-interaction honeypots and runs actual programmable logic [22]; however, it has yet to be used in a study to detect previously unknown ICS attacks, and its hardware emulation may be detectable by a capable attacker [12]. We provide a comparison of several ICS honeypot studies against our own in Section 4.

B. Targeted ICS Attacks

Most, if not all, successful attacks against ICS have been targeted, in that the attackers wish to create adverse physical effects in a specific organisation, and they knowledgeably target specific devices. Examples include Stuxnet attacks against Siemens PLCs [9]; Triton/Trisis attacks against specific models of Schneider Electric's Triconex Safety Instrumented System [7]; and CRASHOVERRIDE attacks against the Ukrainian power grid [10]. The targeted and highly resourced nature of these attacks complicates efforts to identify and track real-world ICS exploitation, as the number of attacks is limited, and attackers have the ability and motivation to limit their exposure. As a result, ICS honeypot studies to date have not identified any attempt to maliciously modify ICS behaviour, nor have they been effectively used to disclose new ICS exploits to the community.

C. Large-scale ICS Attacks

Researchers have demonstrated scalable, proof-of-concept malware for PLCs that modifies programmable logic and automatically spreads to other devices (e.g., to create a botnet or to demand a ransom) [8], [23]. To date, no such large-scale, indiscriminate ICS malware has been observed in the wild. Furthermore, while a decade of security research has demonstrated that tens of thousands of vulnerable ICS devices are directly connected to the internet [21], [24], there has been little evidence of malicious attempts to modify the behaviour of such devices.

The lack of criminal or other large-scale malicious interest in vulnerable ICS devices can be attributed to several economic factors:

- High cost of entry: The cost of hardware for development and testing and the time to gain sufficient knowledge and experience to exploit such devices are significantly higher than in other domains (e.g., IoT).
- Fragmented population: While there may be over 100,000 internet-connected ICS devices, the population is divided amongst dozens of manufacturers running proprietary or bare-metal software on different chipsets.

- Limited resources: ICS devices have limited compute and memory resources, making them poor hosts for resource-intensive tasks such as cryptomining, and they are unlikely to store sensitive information typically used in ransomware attacks. Limited resources and proprietary software make general computing malware unlikely to succeed on ICS devices.

These economic factors are changing as industry seeks new ways to use digital technology. Industry 4.0 and Industrial IoT (IIoT) are converging with the IoT domain [25], creating a larger, more homogeneous environment of low-cost devices with general purpose compute and memory resources. In short, these changes are expected to overcome the economic factors currently inhibiting large-scale malicious interest in the ICS domain. As IIoT and IoT converge and industrial environments become increasingly attractive to cybercriminals and others looking to exploit devices at scale, ICS honeypots will be effective tools to identify and profile these attacks, as they are currently within the IoT domain.

3. SECURIOT DECEPTION TECHNOLOGY

A. ICS Honeypots

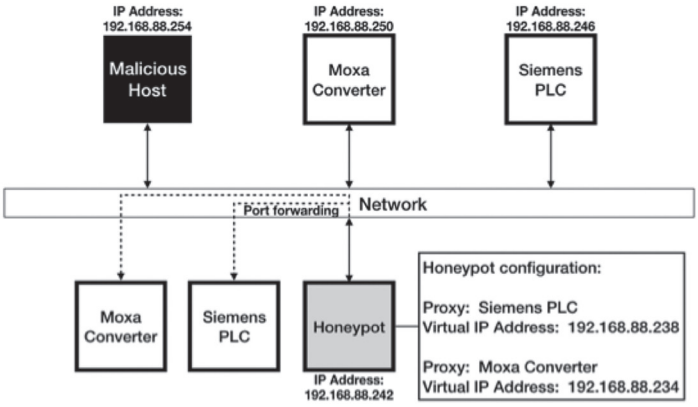
Previous ICS honeypot studies were limited in ways that reduced the likelihood of an attacker being deceived into interacting with the honeypot, such as geographic concentration, the use of cloud hosts, the use of low-interaction honeypots, and short study durations. In this paper, we demonstrate that these limitations can be overcome, showing that a sufficiently-sized, internet-connected ICS honeypot network can be effective in detecting and monitoring previously unknown, targeted attacks.

B. SecuriOT Honeypots

Low-interaction honeypots can be inexpensively deployed at scale, but they are easy to identify. Further, because they do not emulate device state, they cannot be used to profile an attacker's behaviour (e.g., attempts to modify programmable logic). High-interaction honeypots overcome these limitations, but can be expensive to develop, deploy, and maintain. To address the limitations of both low- and high-interaction honeypots, SecuriOT developed a reconfigurable device that supports multiple interaction levels with a common interface and management framework [26]. The device can be configured with templates to emulate an ICS device for low-interaction contexts, like Conpot [14], but can also act as a proxy to a production device. When acting as a proxy, the honeypot redirects traffic to a production device and acts as a man-in-the-middle between the network and the device. This proxy mode allows an adversary to exercise the full behaviour of the target device while providing the honeypot's full logging and alert functionality.

As shown in Figure 1, each physical device is capable of hosting multiple virtual IP addresses and up to three templates simultaneously, allowing a single physical device to appear as multiple devices on a given network.

FIGURE 1: SAMPLE DEPLOYMENT OF A SECURIOT HONEYPOT, SHOWING THE ABILITY TO EMULATE MULTIPLE VIRTUAL DEVICES AND ACT AS A PROXY FOR A CUSTOM DEVICE.



Each physical honeypot interfaces with a Security Information and Event Management (SIEM) system, which logs interactions and raises alerts. Since the honeypot is passive and has no production function on the network, any interaction with a virtual device is suspicious, as it implies that a host is either scanning the network segment or directly interacting with the honeypot. The SIEM is also used to manage device configurations, allowing the honeypots to maintain consistent configurations with the production devices on the network.

C. SecuriOT ICS Honeypot Network

While SecuriOT's ICS honeypots are primarily designed for installation in production systems, the ability to act as a proxy and simultaneously support multiple virtual devices makes them a good foundation for a network of research honeypots. As part of their own intelligence-gathering operation, SecuriOT runs a network of 120 such virtual honeypots with IP addresses geolocated in over 20 countries. Each virtual IP routes traffic to a honeypot acting as a proxy to a production ICS device. Devices include PLCs, RTUs and serial-to-ethernet converters from vendors such as Siemens, Moxa, and Phoenix Contact. The virtual honeypots are supported by up to 15 production devices communicating over the following protocol/port combinations: S7comm/102, BACnet/47808, SOAP/37215, IEC-104/2404, DNP3/20000, and Modbus/502. S7comm, IEC-104, DNP3, and Modbus are used in several industrial environments, including manufacturing, automation, and power and water utilities.

BACnet is used in large-scale building automation. SOAP on port 37215 is used for configuration and management of certain routers.

The honeypots perform full packet captures and SecuriOT performs post-processing, such as fingerprinting the tool used to interact with the honeypot (e.g., NMAP [27]), identifying campaigns, and classifying packets as either reconnaissance or exploitation. The result is a dataset with the fields shown in Table I.

TABLE I: FIELDS PROVIDED PER PACKET FROM SECURIOT'S HONEYPOT NETWORK

Field	Example	Field	Example
Date	2018-03-31	Source country	Japan
Time	06:33:49	Destination country	United States
Source IP address	*[REDACTED]*	Source AS number	AS63949
Source port	51667	Source AS name	Linode, LLC
Destination port	102	Scanning tool	ZMAP
Protocol	S7comm	Campaign	TA-VV
Packet action	Reconnaissance		

4. DATA ANALYSIS AND DISCUSSION

Our dataset consists of 13 months of packets captured between March 2018 and April 2019 from SecuriOT's network of 120, globally-distributed, high-interaction ICS honeypots. The dataset consists of approximately 200,000 packets, which we group into approximately 80,000 interactions. In this section, we present our analysis of the data and discuss our findings. We start with a dataset overview, including a comparison with previous, similar surveys. We then demonstrate malicious use of industrial protocols and discuss the relationships between attackers and targets. We conclude with a demonstration of large-scale attacks against non-industrial protocols recorded by the honeypot network and present early evidence that the ICS domain is affected by malicious, large-scale interest in IoT.

A. Dataset Overview

Table II provides a summary of the interactions with the SecuriOT network of industrial honeypots over the period of observation. The data demonstrates the breadth of interest in internet-connected ICS devices: thousands of individual hosts (IP addresses) are scanning industrial protocols from dozens of Autonomous Systems (ASes) in dozens of countries.

We find that a majority of these interactions originate from well-known research scanners and are expected to be benign (e.g., Censys [28], Shodan [29]), which is consistent with previous observations [19] – [21]. Both SecuriOT and the Cambridge Cybercrime Centre (CCCC) [30] maintain lists of known scanners, against which source IP addresses were compared to generate the 'Known scanners' percentages in Table II. Similarly, Table II shows that a vast majority of interactions are initiated by well-known scanning tools, such as NMAP [27].

Following previous studies, we classify multiple received packets from a given IP address as part of a single 'interaction'. Comparing interactions rather than packets is preferable because the number of packets required to perform a given task can vary for different scanning tools and protocols. We define an interaction as a single scanning or exploitation event. For example, the Siemens module from the ZGrab scanner sends about 12 packets to each scanned IP address, while scanning a single port with ZMap only sends two packets (TCP SYN and RST) to each scanned IP address [21], [31]. Each of these would be considered one interaction.

TABLE II: SUMMARY OF INTERACTIONS WITH SECURIOT'S NETWORK OF INDUSTRIAL HONEYPOTS

Protocol/Port	Total packets	Related Interactions	Source IP addresses	Source ASes	Source countries	Known scanners	Known tools
Modbus/502	54,682	18,980	1,321	91	31	69.5%	99.9%
BACnet/47808	50,276	20,097	1,073	35	16	84.7%	100.0%
S7comm/102	43,203	18,422	998	85	30	49.9%	99.5%
DNP3/20000	32,534	13,283	1,040	124	42	39.1%	99.9%
SOAP/37215	12,975	7,403	337	85	29	0.0%	51.2%
IEC-104/2404	8,797	3,404	214	162	23	7.0%	99.6%

B. Comparison with Earlier Studies

Different studies use different methodologies and focus on different protocols; therefore, direct comparison is challenging. Even surveys covering the same timeframe but using different methodologies can produce different results (e.g., network telescopes versus honeypots [21]). We approach such comparisons with caution, and only draw qualitative conclusions. We selected studies for comparison for the following reasons: Mirian *et al*. is regularly used for comparison in other studies [21]; Ferretti *et al*. is a more recent study of similar size to Mirian *et al*. and has a global scope [19]; and Cabana *et al*. is the largest, low-interaction ICS honeypot study in the literature [20]. Notably, all three of these studies use low-interaction honeypots, whereas our

study uses high-interaction honeypots. There is no comparable survey in the academic literature of a large-scale, high-interaction ICS honeypot network.

Table III compares these surveys and data collection methods, showing broad agreement in the observed scanning frequency against each protocol. Table III also demonstrates that ranking based on interactions results in a different ordering than ranking based on packets, as different protocols have different packet densities.

TABLE III: COMPARISON OF RANKED POPULARITY OF SCANNED INDUSTRIAL PROTOCOLS FROM MULTIPLE SURVEYS.
'*' indicates raw data was not available. '**' indicates an estimate based on graphical data.

Source	Method	Dataset type	Dataset size	Ranked popularity					
SecuriOT	Honeypot	Packets	202,467	Modbus	BACnet	S7comm	DNP3	IEC-104	
SecuriOT	Honeypot	Interactions	81,589	BACnet	Modbus	S7comm	DNP3	IEC-104	
Mirian et al. [21]	Telescope	Packets	2,100	Modbus	BACnet	S7comm	DNP3	Ethernet/IP	
Mirian et al. [21]	Honeypot	Interactions	5,252	S7comm	Modbus	BACnet			
Ferretti et al. [19]	Honeypot	Packets	*	Modbus	BACnet	S7comm	Ethernet/IP	IEC-104	
Ferretti et al. [19]	Honeypot	Interactions	4,986	BACnet	Modbus	Ethernet/IP	S7comm	IEC-104	
Cabana et al. [20]	Telescope	Packets	197M**	BACnet	Modbus	DNP3	S7comm	Ethernet/IP	IEC-104

While the distribution of our scanning traffic is largely consistent with previous studies, the data shows both growth and asymmetry in the DNP3 scanning traffic that has not been previously identified or evaluated. Mirian *et al.* only identified 5.1% of network telescope traffic as targeting DNP3 in 2015 [21], whereas Cabana *et al.* observed over 22% of network telescope data targeting DNP3 in 2019 [20]. Similarly, over 16% of the interactions recorded by SecuriOT honeypots targeted the DNP3 protocol. Furthermore, as shown in Table II, while the total number of DNP3 interactions is only 70% of the number of Modbus interactions (13,283 vs. 18,980), the number of IP addresses scanning for DNP3 is nearly 80% of that of Modbus (1,040 vs. 1,321), and the number of ASes from which those IP addresses originate is 136% of those for Modbus (124 vs 91). This statistic is also reflected in the number of source countries in which those IP addresses are geolocated (42 vs. 31). The asymmetry is even more pronounced when comparing DNP3 with BACnet or S7comm. Despite the challenges in quantitative comparisons between studies, there is clear evidence from multiple studies demonstrating a wider, as well as a growing, interest in DNP3 compared to other industrial protocols.

C. Targeted Attacks via Industrial Protocols

SecuriOT's analysis concludes that only 20 of the 200,000 captured packets make use of an industrial protocol with clear malicious intent. These 20 packets can be grouped into nine attack interactions, which are summarised in Table IV. Based on feedback

from vendors and vulnerability databases, four of the nine interactions represent previously unknown attacks, or zero days, and one represents the first documentation of a previously-identified proof-of-concept attack in the wild [32]. The attack types include denial of service (DoS) and command replay attacks.

TABLE IV: ATTACKS USING INDUSTRIAL PROTOCOLS. THE PACKET COUNT DOES NOT INCLUDE TRANSPORT LAYER HANDSHAKES (E.G., INITIAL SYN PACKET FOR PROTOCOLS LAYERED ON TCP).

Date	Source country	Destination country	Protocol	Attack type	Source AS number	Number of packets
2 Apr 2018	United States	China	IEC-104	DoS	AS394828	2
17 Apr 2018	China	Poland	IEC-104	DoS	AS4134	1
20 Apr 2018	Russia	United States	S7comm	Replay	AS60307	8
27 Jun 2018	Ukraine	China	IEC-104	DoS	AS15626	4
8 Aug 2018	Vietnam	France	S7comm	DoS	AS38731	1
8 Aug 2018	Vietnam	Lithuania	S7comm	DoS	AS38731	1
9 Aug 2018	Vietnam	Poland	Modbus	DoS	AS38731	1
9 Aug 2018	Vietnam	France	Modbus	DoS	AS38731	1
19 Nov 2018	Seychelles	Czech Republic	Modbus	DoS	AS29073	1

The DoS attacks took several forms. In one case, a specially crafted packet forced a device to violate its real-time constraints, providing a low-bandwidth DoS attack on the process control. In another case, the attack targeted devices with incomplete implementations of the protocol stack; the attack provided valid, but unimplemented commands, and adversely affected the device's process control. The attacker specifically targeted vulnerable device types, so this was not a case of accidental DoS. In a third case, a buffer overflow affected the device's network communication capability, but did not affect the device's process control.

Since many industrial protocols lack authentication or encryption, the receipt of any packet with a parsable command may be considered valid. In some cases, though, manufacturers have implemented protection to prevent a replay of previous commands or commands recorded in a test environment. The replay attack identified by SecuriOT was successful against a device for which the manufacturer claimed replay protection.

For most of the attacks, the source IP address was only active for the attack itself; the honeypot network had no record of other interactions from that IP address. This is not unexpected: an attacker may use one or multiple IP addresses for reconnaissance and then use a fresh IP address for the actual attack, to avoid blacklists. For three of

the attacks, however, consistent activity was observed from the source IP address. Specifically, the IP addresses used for the attacks originating in Vietnam, Ukraine, and the Seychelles performed regular scanning over the entire study duration.

These vulnerabilities and associated exploits were responsibly disclosed by SecuriOT to the device manufacturers, and public disclosure is currently being negotiated. The relationship between these vendors and SecuriOT precludes further public disclosure of the vulnerabilities at this time, but additional details may be obtained in some cases directly from SecuriOT.

D. Large-scale Attacks via Industrial Protocols

SecuriOT's honeypots also exposed a non-industrial protocol port and captured data associated with the Okiru-Satori variant of the Mirai botnet, which is indiscriminate and targets any vulnerable device across any network to which an infected device is connected.

While Okiru-Satori does not target industrial protocols, the convergence of IIoT and IoT domains may result in industrial devices being included in large-scale, non-industrial attacks. This is already the case for Windows-based industrial infrastructure. For example, the ransomware attack against the Windows-based infrastructure at Norsk Hydro in early 2019 prevented the safe and effective use of industrial devices [33]. As IIoT devices incorporate common operating systems with general purpose processing (e.g., Linux-based Azure Sphere [34]), they are more likely to become inadvertent victims of large-scale botnet or ransomware attacks targeting the IoT population. In this section, we discuss interactions with Mirai hosts and show that overlap already exists with industrial protocol scanners.

The Mirai botnet emerged in 2016 and used aggressive scanning and brute force password searches to infect hundreds of thousands of Linux-based IoT devices. At its peak, an estimated 600,000 hosts were infected [35]. At the time of writing, the CCCC [30] observes approximately 150,000 infected hosts per day scanning IP addresses in a monitored /14 network. The scanning packet used by Mirai is distinctive, allowing the CCCC to identify suspected Mirai hosts and record data such as the source and destination IP addresses and port numbers.

Many Mirai variants emerged after the public release of the Mirai source code. Variants target different device types and architectures and exploit different vulnerabilities. The Okiru-Satori variant was identified in 2017 and targeted Huawei routers on port 37215 using a previously unidentified vulnerability (CVE-2017-17215) [36]. As shown in Table V, SecuriOT's honeypots recorded 7,403 interactions from 337 IP addresses on port 37215. Of these, SecuriOT identified 222 malicious interactions, based on

attempts to brute force passwords, make use of the vulnerabilities exploited by Okiru-Satori, or modify firmware. While the malicious packets make up more than 30% of the total traffic on port 37215, only 3.0% of the total interactions are malicious, as password searches and firmware downloads necessarily require more packets than scanning.

TABLE V: SUMMARY OF INTERACTIONS ON PORT 37215.

	Packets	Interactions	Source IP Addresses	Source ASes	Dates of interaction
Overall	12,975	7,403	337	85	266
Malicious	3,919	222	13	2	15

Notably, while the scanning of port 37215 was recorded on 266 days, the honeypots were only configured as vulnerable routers over short periods in April and July 2018, resulting in only 15 days of malicious interactions. As discussed below, some of the apparently benign scanning might have transitioned to exploitation had the scanner found the honeypot in a vulnerable configuration.

To study the overlap between Mirai hosts and hosts aware of industrial protocols, we combined the CCCC database of suspected Mirai hosts [30] with SecuriOT's honeypot data, correlating source IP addresses and interaction dates. Table VI summarises the results of this comparison from the perspective of the SecuriOT honeypots. For example, the first row should be interpreted as 792 packets received by SecuriOT honeypots on port 37215 from 26 IP addresses that the CCCC suspected to be hosting Mirai on the day of the interaction with the honeypot.

TABLE VI: SECURIOT HONEYPOT DATA CORRESPONDING TO SOURCE IP ADDRESSES AND DATES FROM THE CCCC MIRAI HOST DATASET.

Protocol/Port	Total packets	Related Interactions	Source IP addresses	Source ASes	Dates
SOAP/37215	792	789	26	11	40
DNP3/20000	116	71	4	2	4
BACnet/47808	2	2	1	1	1
Modbus/502	1	1	1	1	1

Comparing 789 SOAP/37215 interactions in Table VI with 222 malicious interactions in Table V demonstrates that the CCCC suspects many of the benign interactions with SecuriOT honeypots to have originated from Mirai hosts that simply did not find the SecuriOT honeypot to be vulnerable. This is consistent with the knowledge that the SecuriOT honeypots were only configured as vulnerable routers during limited periods.

Table VI also shows that the CCCC suspects 74 industrial protocol interactions (i.e., over DNP3, BACnet and Modbus) with SecuriOT honeypots to have originated from IP addresses hosting Mirai. As there is no known variant of Mirai that targets ICS devices, the scanning traffic by Mirai hosts against industrial protocols implies either that these scanners share an IP address with a Mirai host (e.g., a scanner behind an infected router) or that the scanner uses a similar technique to that employed by Mirai, though we are not aware of any such benign, internet-wide scanners.

This overlap between SecuriOT's honeypot data and the CCCC Mirai database, though limited, suggests that the gap between ICS-aware and IoT-aware hosts is narrowing.

5. RECOMMENDATIONS FOR HONEYPOT NETWORKS

SecuriOT's honeypot network exposed four zero-day attacks against devices running common ICS protocols, such as S7comm and Modbus. By comparing our study with previous studies that did not identify similar exploits (e.g., [19] – [21]), we provide the following recommendations for deploying networks of ICS honeypots for security research:

- Honeypot networks should be geographically dispersed. We identified nine attacks against devices in six countries, and none of the attacks originated in the same country as the target. Several honeypot studies located most or all targets in the United States [19], [21]; however, of our nine identified attacks, only one target was located in the United States.
- Honeypots should be hosted at realistic IP addresses. Several previous ICS honeypot studies used AWS or other cloud providers to host honeypots [19] – [21]. ICS devices are unlikely to be connected via a cloud service provider, so the use of AWS or similar is a red flag to an attacker.
- Honeypots should be high-interaction. Low-interaction honeypots can often be fingerprinted and generally do not allow an attacker to interact with the device beyond the initial login screen or protocol handshake. To deceive targeted attackers and understand their intentions (e.g., modifying firmware or programmable logic), high-interaction honeypots are necessary.

- Honeypot use should be systematic and continuous. This provides both authenticity and a larger window for an attacker to identify and target a given honeypot. Unlike large-scale attacks scanning for any vulnerable device, targeted attackers are looking for specific devices and may take considerable time before accidentally targeting a honeypot.

6. CONCLUSIONS

We have demonstrated that a network of high-interaction honeypots can identify and profile previously unknown, targeted ICS attacks. Specifically, we exposed four zero-day attacks against devices running common ICS protocols such as S7comm and Modbus, which were disclosed to the applicable manufacturers.

We also demonstrated that the gap between ICS-aware and IoT-aware hosts is narrowing, showing that IoT malware is co-located with ICS devices and scanners. Bridging this gap is the first major hurdle in attacking ICS devices at scale. Thus far, ICS devices have not been subjected to indiscriminate targeting, but the convergence of IIoT and IoT domains will make industrial devices more attractive targets, even if only as a vulnerable sub-population amongst the growing IoT population.

Finally, we discussed the limitations of previous ICS honeypot studies and provided recommendations for developing effective ICS honeypot networks as intelligence-gathering tools.

ACKNOWLEDGEMENTS

Michael Dodson is supported by a scholarship from the Gates Cambridge Trust; Alastair R. Beresford is partially supported by EPSRC [grant number EP/M020320/1]. The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect those of any of the funders. We thank Alexander Vetterl for his feedback and patient sharing of his expertise.

REFERENCES

[1] B. Seri, G. Vishnepolsky, and D. Zusman, "URGENT/11 technical white paper," *Armis* [Online]. Available: https://go.armis.com/urgent11. [Accessed: 25-Nov-2019].
[2] A. Nochvay, "Security research: CODESYS runtime, a PLC control framework," Kaspersky ICS CERT, 2019 [Online]. Available: https://perma.cc/325P-N7AV. [Accessed: 08-Nov-2019].
[3] Cybersecurity and Infrastructure Security Agency, (2013) "3S CoDeSys vulnerabilities," [Online]. Available: https://perma.cc/F8W4-7H75. [Accessed: 28-Oct-2019].

[4] O. Karliner, “FreeRTOS TCP/IP stack vulnerabilities,” *Zimperium Mobile Security Blog*, 4 December 2018. [Online]. Available: https://blog.zimperium.com/freertos-tcpip-stack-vulnerabilities-details/. [Accessed: 18-Dec-2019].
[5] C. Mihalcik, “Microsoft aims for 1 billion devices running Windows 10,” *CNET*. [Online]. Available: https://www.cnet.com/news/microsoft-aims-for-1-billion-devices-running-windows-10/. [Accessed: 26-Feb-2020].
[6] É. Leverett, R. Clayton, and R. Anderson, “Standardisation and certification of the ‘Internet of Things’,” Workshop on the Economics of Information Security (WEIS), 2017 [Online]. Available: https://perma.cc/5Y9R-9DD3.
[7] Dragos Inc. “TRISIS malware: Analysis of safety system targeted malware”. [Online]. Available: https://perma.cc/K9EM-CABV. [Accessed: 08-Nov-2019].
[8] D. Formby, S. Durbha, and R. Beyah, “Out of control: Ransomware for industrial control systems,” RSA Conference, 2017 [Online]. Available: https://perma.cc/XD8V-LP5M.
[9] T. Chen and S. Abu-Nimeh, “Lessons from Stuxnet,” *Computer*, vol. 44, no. 4, 2011, doi: 10.1109/MC.2011.115. [Online]. Available: http://ieeexplore.ieee.org/document/5742014/. [Accessed: 14-Jun-2019]
[10] R. M. Lee, “CRASHOVERRIDE: Analyzing the malware that attacks power grids,” *Dragos Inc*., 2017 [Online]. Available: https://dragos.com/resource/crashoverride-analyzing-the-malware-that-attacks-power-grids/. [Accessed: 19-Jun-2019].
[11] Cybersecurity and Infrastructure Security Agency, “Russian Government Cyber Activity Targeting Energy and Other Critical Infrastructure Sectors,” [Online]. Available: https://www.us-cert.gov/ncas/alerts/TA18-074A. [Accessed: 26-Feb-2020].
[12] A. Vetterl and R. Clayton, “Honware: A virtual honeypot framework for capturing CPE and IoT zero days,” Symposium on Electronic Crime Research (eCrime), Pittsburgh, United States of America, 2019.
[13] W. Martin, “Honey pots and honey nets - security through deception,” *SANS Institute*, 2003 [Online]. Available: https://www.sans.org/reading-room/whitepapers/attacking/honey-pots-honey-nets-security-deception-41. [Accessed: 08-Apr-2019].
[14] L. Rift, J. Vastergaard, D. Haslinger, A. Pasquale, and J. Smith, CONPOT ICS/SCADA honeypot.. [Online]. Available: http://conpot.org. [Accessed: 08-Apr-2019].
[15] R. Rustici and I. Barak, “ICS threat broadens: Nation-state hackers are no longer the only game in town,” *Cybereason*. [Online]. Available: https://www.cybereason.com/blog/industrial-control-system-specialized-hackers. [Accessed: 09-Dec-2019].
[16] L. Spitzner, “The value of honeypots, part one: Definitions and values of honeypots,” *Symantec*, 2001. [Online]. Available: https://www.symantec.com/connect/articles/value-honeypots-part-one-definitions-and-values-honeypots. [Accessed: 09-Dec-2019].
[17] N. Provos and T. Holz, *Virtual honeypots: From botnet tracking to intrusion detection*. Pearson Education, 2007.
[18] W. Fan, Z. Du, D. Fernández, and V. A. Villagrá, “Enabling an anatomic view to investigate honeypot systems: A survey,” *IEEE Systems Journal*, vol. 12, no. 4, 2018, doi: 10.1109/JSYST.2017.2762161.
[19] P. Ferretti, M. Pogliani, and S. Zanero, “Characterizing background noise in ICS traffic through a set of low interaction honeypots,” *ACM Workshop on Cyber-Physical Systems Security & Privacy (CPS-SPC)*, 2019, doi: 10.1145/3338499.3357361.
[20] O. Cabana, A. M. Youssef, M. Debbabi, B. Lebel, M. Kassouf, and B. L. Agba, “Detecting, fingerprinting and tracking reconnaissance campaigns targeting industrial control systems,” International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, 2019, doi: 10.1007/978-3-030-22038-9_5. [Online]. Available: http://link.springer.com/10.1007/978-3-030-22038-9_5. [Accessed: 09-Aug-2019].
[21] A. Mirian *et al*., “An Internet-wide view of ICS devices,” *Conference on Privacy, Security and Trust (PST)*, Auckland, New Zealand, 2016, doi: 10.1109/PST.2016.7906943.
[22] D. Antonioli, A. Agrawal, and N. O. Tippenhauer, “Towards high-interaction virtual ICS honeypots-in-a-box.” presented at ACM Workshop on Cyber-Physical Systems Security and Privacy (CPS-SPC)] Vienna, Austria, 2016, doi: 10.1145/2994487.2994493. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2994487.2994493. [Accessed: 20-Dec-2019].
[23] R. Spenneberg, M. Brüggemann, and H. Schwartke, “PLC-Blaster: A worm living solely in the PLC,” presented at Black Hat Asia Marina Bay Sands, Singapore, 2016 [Online]. Available: https://perma.cc/XWU5-TZ7L. [Accessed: 28-Oct-2019].
[24] É. P. Leverett, “Quantitatively assessing and visualising industrial system attack surfaces,” *University of Cambridge MPhil Thesis*, 2011 [Online]. Available: https://perma.cc/83Z9-Q5J9. [Accessed: 26-Feb-2019].

[25] M. Wollschlaeger, T. Sauter, and J. Jasperneite, "The future of industrial communication: Automation networks in the era of the internet of things and Industry 4.0," *IEEE Industrial Electronics Magazine*, 2017, doi: 10.1109/MIE.2017.2649104. [Online]. Available: http://ieeexplore.ieee.org/document/7883994/. [Accessed: 28-Oct-2019].
[26] SecuriOT, "SecuriOT honeypot: Powered by Industrial Defenica," *SecuriOT Honeypot - Powered by Industrial Defenica*. [Online]. Available: https://www.honeypot.dk. [Accessed: 28-Dec-2019].
[27] NMap Project, "Nmap: the network mapper," *NMap Project*. [Online]. Available: https://nmap.org/. [Accessed: 02-Dec-2019].
[28] "Censys," *Censys*. [Online]. Available: https://censys.io/. [Accessed: 28-Dec-2019].
[29] "Shodan," *Shodan*. [Online]. Available: https://www.shodan.io/. [Accessed: 28-Dec-2019].
[30] Cambridge Cybercrime Centre, "Computer Laboratory: Cambridge Cybercrime Centre: Description of available datasets," Cambridge Cybercrime Centre [Online]. Available: https://www.cambridgecybercrime.uk/datasets.html. [Accessed: 01-May-2019].
[31] "The ZMap project," *ZMap Project*. [Online]. Available: https://zmap.io/. [Accessed: 02-Dec-2019].
[32] Cybersecurity and Infrastructure Security Agency, "PLC cycle time influences (Update A)". [Online]. Available: https://www.us-cert.gov/ics/advisories/ICSA-19-106-03. [Accessed: 29-Feb-2020].
[33] Norsk Hydro, "Cyber-attack on Hydro," Norsk Hydro ASA, Oslo, Norway, 2019[Online]. Available: https://www.hydro.com/en/media/on-the-agenda/cyber-attack/. [Accessed: 14-Dec-2019].
[34] Microsoft Azure, "Azure Sphere," Microsoft Corporation, Redmond, WA,. [Online]. Available: https://azure.microsoft.com/en-us/services/azure-sphere/. [Accessed: 13-Jun-2019].
[35] M. Antonakakis *et al*., "Understanding the Mirai botnet," USENIX Security Symposium (USENIX Security), Vancouver, Canada, 2017 [Online]. https://www.usenix.org/system/files/conference/usenixsecurity17/sec17-antonakakis.pdf. [Accessed: 10-Jun-2019].
[36] Check Point Research, "Huawei home routers in botnet recruitment," Check Point Software Technologies Inc., San Carlos, CA , Dec. 2017 [Online]. Available: https://research.checkpoint.com/2017/good-zero-day-skiddie/. [Accessed: 30-Nov-2019].

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Addressing the Cybersecurity Challenges of Electrical Power Systems of the Future

Gilberto Pires de Azevedo
Researcher
Electrical Power Systems Dept.
Electric Energy Research Centre
Rio de Janeiro – RJ, Brazil
gilberto@cepel.br

Maxli Barroso Campos
Chief of Division
Strategic Management Dept.
Cyber Defence Command
Brasília – DF, Brazil
campos@cdciber.eb.mil.br

Paulo César Pellanda
Professor
Electrical Engineering Dept.
Military Institute of Engineering
Rio de Janeiro – RJ, Brazil
pcpellanda@ieee.org

Abstract: Electrical Power Systems (EPSs) are among the most prominent critical infrastructures of our digital society. Assets, systems and networks of most other critical infrastructure sectors depend heavily on EPSs and would fail in the event of persistent electricity supply problems. This should make EPSs attractive targets for cyberattacks, so it is somewhat surprising that few large-scale successful cyberattacks on the electricity sector have been reported so far.

EPSs structures are undergoing deep changes that will accelerate over the next years. A convergence of environmental concerns and technological evolution is leading to the widespread use of distributed renewable microgeneration, electric vehicles, distributed energy storage, Internet of Things, smart grids and software-defined operating devices. These game-changing innovations are reshaping EPSs. The previously well-ordered computational environment where a limited number of agents interacted in

predictable ways will gradually receive new layers of agents, where thousands or even millions of them will buy or sell services in a kind of giant open market. The search for individual advantages or profits rather than overall system welfare will guide the actions of these new participants.

This work examines the traditional structure of EPSs from a cybersecurity point of view as well as foreseeable changes. It will also look at associated risks and discuss possible approaches to mitigate them.

Keywords: *critical infrastructure, electrical power systems, soft cybersecurity, industrial control systems, SCADA*

1. INTRODUCTION

Discussions about cybersecurity concerns in electrical power systems (EPSs) quite often have an alarmist approach. Catastrophic scenarios in which cyberattacks produce massive blackouts leading to generalised chaos and substantial economic losses are imagined and described by specialists trying to draw attention to cybersecurity issues in EPSs. However, there are as yet no examples of cyberattacks that have had such drastic consequences, which can lead some sceptical decision-makers to neglect prevention.

It can be said that both views are right to a point. While the possibility of cyberattacks with catastrophic consequences remains small today, it will increase quickly over the 2020s, mainly on the back of the profound transformations taking place in the electricity supply sector. Satisfactory cybersecurity levels are not only a condition for the safe operation of systems but also a crucial requirement for system evolution. Preventive measures to mitigate the vulnerabilities and risks of the new environment are possible and essential, but significant work on research, development, governance and other areas is required to provide and maintain acceptable levels of cybersecurity.

This work starts with an overview of the evolution of EPSs from the perspective of cybersecurity (Sections 2 and 3), followed by a discussion of some foreseeable challenges (Section 4). Possible approaches to tackle those challenges and suggestions for future work are presented in Section 5.

For the sake of conciseness, the abbreviation EPS stands for “electrical power system”, comprising generation, transmission and distribution equipment and, in

some cases, also the associated computational and communication infrastructure. For the same purpose, the text avoids addressing general cybersecurity concepts except when necessary to examine specific details about EPSs. The term "attack" (and hence "attacker" and "cyberattack") is used throughout the text in a broader and more informal sense than defined in [1]. Finally, since multi-agent systems are an appropriate metaphor to represent EPSs of the future, the term "agent" is applied to refer to any active participant of the system that has some degree of autonomy for monitoring the environment, communicating with some other agents and acting to reach its own goals [2, 3].

2. ELECTRICAL POWER SYSTEMS: CRITICAL INFRASTRUCTURE FOR OUR SOCIETY

EPSs are among the most prominent critical infrastructures of our digital society. The assets, systems and networks of most other critical infrastructure sectors depend heavily on EPSs and would fail in the event of persistent electricity supply problems, generating a ripple effect and seriously compromising other critical infrastructures [4].

This should make EPSs very attractive targets for cyberattacks; thus it is somewhat surprising that few successful large-scale attacks have been reported in the electricity sector so far. Nevertheless, a closer look shows that the vulnerability of existing electric power grids to cyberattacks is not too alarming at present, in part due to the relative abundance of old elements with low degrees of computational connectivity, as well as the still small number of different classes of agents that interact via computer networks.

A complementary explanation for the relative success of cyber protection of EPSs today is the still limited motivations for cyberattacks – especially the scant possibility of obtaining economic advantages from them. Unlike from attacks on services such as banking, there are as yet few possible rewards to be gained from attacking EPSs.

As an illustration, one can examine the famous December 23, 2015 cyberattack at Ukrainian Kyivoblenergo [5], a regional electricity distribution company. This incident is often reported as an example of the potential effects of attacks on EPSs. Very sophisticated techniques were applied, and months of preparation were required. Up to 225,000 customers were affected by power outages that lasted several hours; however, the impacts of the incidents were rated as low, as the outages affected a small number of overall power consumers in Ukraine and were limited in duration. Analysis results based on a single incident should not be generalised, but the balance

between the likely effort expended in preparing that attack and its results does not seem to encourage further similar attacks.

Unfortunately, this relatively peaceful scenario will not last for long. EPSs are undergoing profound structural changes that will make cybersecurity a primary concern for system regulators, planners and operators [6] – and not only for them.

3. ELECTRICAL POWER SYSTEMS: DEEP TRANSFORMATIONS TAKING PLACE

EPSs are perhaps the most extensive and complex artificial infrastructures on Earth, but have been evolving slowly and incrementally for decades. Despite changes in governance in some countries, the physical structure of EPSs has remained essentially the same for a long time. Utilities, consumers, regulators and operators have well-defined roles and interact in a well-ordered fashion. Computational systems and communication networks associated with EPS monitoring and control are often isolated from other networks and based on non-standard implementations. Cybersecurity preventive measures are incipient, but prospective cyber attackers have had a small surface of attack available and the possible consequences of successful attacks have tended to be limited in extension and duration.

However, EPS structures are currently undergoing profound changes that will accelerate in the coming years. A convergence of environmental concerns, technological evolution and other drivers will reshape EPSs over the next decades:

a. The uncertainty in the availability of generation due to the widespread use of intermittent distributed renewable generation like wind and photovoltaic;
b. Expected advances in distributed electricity storage technology;
c. Electric vehicles that might behave either as moving loads or electricity storage devices;
d. New roles for consumers, who will gradually change their passive behaviour to act also as small energy producers and energy stores; they will also be able to autonomously control their demand in response to dynamic energy prices or similar indications;
e. Internet of Things, 5G and other innovations will connect vast numbers of sensors and control devices to EPSs. Even some domestic apparatus will be connected and respond with a certain degree of autonomy to external signs and demands.

The drivers behind these transformations in EPSs are often grouped under the so-called "3-Ds" view: digitalisation, decarbonisation and decentralisation. Smart grid, autonomic power systems [3] and multi-agent systems [7] are concepts that provide abstractions that help to handle the complexity of the future EPS environment [9].

A. New Layers of Agents

Long-established EPS actors like utilities, customers, operators, regulators and similar ones [9] could be classified as the "first layer" of agents; the "second layer" would encompass new classes of agents that are just starting to take part in the electrical power system such as distributed microgenerators, electric vehicles and storage units [9]; "third layer" agents would include, among others, associations of agents of previous layers; and the "fourth layer" includes providers of services for associations of agents, etc. The resulting environment will be diversified, probably following this proposal for stratification in different layers, with a number of agents far greater than that of the existing "first layer". The previously well-ordered computational environment where a limited number of agents interacted in predictable ways will coexist with – or be replaced by – a much more complex one where a vast number of agents will buy or sell services in a kind of giant open market. In [3], the author mentions "the potential for hundreds of millions of devices across Europe to be involved in the electricity market and to contribute to network operation through demand response" by 2050. The search for individual advantages or profits rather than overall system welfare will guide the actions of most of these new agents.

By the early 1990s, when the internet took its first steps outside research institutions, it was already clear to many that it was a habitat where a plethora of new businesses would emerge and evolve in a very different way than in the physical world. However, despite a handful of evident candidates (news, banking, marketing, commerce and a few others), at that time no one could have predicted the extraordinary diversity of new businesses that would appear on the internet, nor the associated risks. Electrical power system researchers and planners are currently in a situation that resembles that of internet pioneers: while it is evident that many new businesses and agents will start to have active roles in the system in the coming years, it is challenging to guess precisely who they will be and how tightly controlled the environment where they will interact will be.

In short, the expected transformations suggest that EPS cybersecurity professionals will have to deal with an increase in both cyber vulnerabilities and attack surfaces, with widespread connections to potentially insecure external networks, and with explosive numbers of new and relatively independent active agents.

B. Increasing Criticality of SCADA Systems

SCADA ("supervisory control and data acquisition") systems are often part of industrial control systems (ICS) that monitor and control industrial processes. Few of these processes are as relevant to our society as EPS control, where SCADA systems are the main actors. Due to their criticality, they deserve special attention in any cybersecurity analysis.

Early generations of SCADA systems were built over proprietary technologies and often used customised versions of communication protocols; connections to the internet were rare. Although cyber protection measures were almost non-existent, those SCADA systems were relatively protected from cyberattacks by a combination of "security through obscurity", small surface of attack and limited motivations for cyber attackers.

As mentioned earlier, this peaceful landscape is changing rapidly. SCADA systems are now directly or indirectly connected to the internet, use standard communication protocols, and proprietary technologies have been replaced by commercial software packages and operational systems. Despite providing significant reductions on development and evolution costs and schedules, improving maintainability and favouring interoperability, in theory these changes could make SCADA systems increasingly vulnerable to even generic malware attacks. Adding to this scenario the increased motivations for attackers, SCADA systems will face significant cybersecurity challenges.

Frameworks like the Purdue Model for Control Hierarchy [8] provide good starting points for the segmentation of EPS control systems, including SCADA, and help to build more secure environments by defining zones with different protection requirements. It is likely that such frameworks will need to be expanded to cover the interactions of SCADA with some of the agents of layers two through four mentioned previously. Interactions with them will significantly differ from others like those with process devices or elements in corporate networks, thus demanding the definition of specific security requirements.

The specificities of the cybersecurity of SCADA environments, discussed below, are sometimes not well understood by professionals of other areas to which those systems are now connected, such as corporate networks. The "availability-first" approach, whereby service continuity is far more important than data confidentiality or even data integrity, may clash with corporate cybersecurity policies.

4. CYBERSECURITY CHALLENGES OF THE EPSs OF THE FUTURE

A. Cyberattacks: Motivations and Targets

It should be noted that there are no significant difficulties in making successful low-tech physical attacks on the electricity grid. Transmission facilities, for example, can be dropped down with simple tools, and simultaneous coordinated attacks on a few strategic transmission lines can lead to severe and long-term outages. The rarity of such attacks suggests that, in peacetime, there are not many motivations for triggering broad and unfocused power shutdowns.

However, in EPSs of the future, increasing cyber vulnerabilities, attack surfaces and severity of effects, and the feasibility of remote attacks without immediate risk to attackers, are likely to reinforce the motivations of cyberattacks. War, terrorism, vandalism and different brands of radical activism are some ordinary motivations for cyberattacks that could be aimed at causing large-scale electricity shutdowns. Other motivations related to criminal activities might also gain relevance. The extortion of power utility companies through threats of cyberattacks that could cause power outages is another example of a set of new options that cybercriminals might try to exploit; new successful criminal "business models" can appear at any time.

Advances in smart grid, Internet of Things and digitalisation in general are opening doors to sharply focused attacks with a renewed set of motivations such as revenge, privacy breaching, harming business competitors and cyber versions of ordinary crimes. For instance, a hacker may try to remotely turn off the heating system of his ex-girlfriend's home, shut down the electricity of an obnoxious neighbour, produce overvoltage to damage equipment of a competitor, or steal credits from microgenerators. Such focused cyberattacks can become very common if insufficient preventive measures are taken.

On the other hand, as mentioned before, advances in the use of commercial software on SCADA systems and other EPS control systems can make them vulnerable to generic malware attacks with motivations that are not related to EPSs.

B. Beyond Cyberattacks

The new EPS scenario described in the previous section, besides bringing new motivations and opportunities for cyber attackers, adds myriad agents that could hardly be called "attackers" but may behave in ways that would harm other agents or even the whole system [9]. A few examples are:

a. Formerly well-behaved agents that are facing temporary problems and thus unable to respond appropriately to requests from other agents, or have had their behaviour degraded permanently whether intentionally or not;
b. Rogue agents offering services that they are unable to provide adequately, due to quantity, quality, or timing issues;
c. Agents trying to mislead their customers to increase their revenues;
d. Agents acting to harm competitors using unfair methods;
e. Agents trying to obtain advantages or revenues illegally.

There will likely be other examples of ill-behaved agents in the EPS of the future. This situation may be a novelty for EPS professionals accustomed to well-controlled computational environments, but not for internet professionals familiar with the risks of open environments. Soft cybersecurity metrics like trust and reputation can help in such environments, as will be seen later.

C. Cyber Operations Against EPSs

Despite fortunately being one of the least common kinds of attacks, cyber operations [1] against EPSs are serious concerns and require a wide range of defensive measures (offensive actions are not discussed in this work). Such operations are likely to be conducted by terrorists, military personnel or sectors of a foreign government. An operational target might be a set of critical cyber infrastructures that include EPSs, an EPS itself, or a more specific objective, such as part of an EPS that feeds power to a specific city, industry or military facility. Sections of an EPS that supply power to military command and control facilities or to weapon systems are also among some preferential targets. Unlike during the Cold War when there were "demonstrations" of the effect of new military technologies, so far cyber operations have tended to be apocryphal [10].

The growing interdependence between critical infrastructures – such as EPSs and communication networks – increases vulnerabilities and the complexity of cyber defence planning. Since technical, practical and economic reasons make it impossible to guarantee comprehensive protections for all critical infrastructures against all threats and risks, identifying key vulnerabilities and infrastructures and critical points to be protected is essential [11].

The evolution of EPSs requires a specialised treatment to identify new intra- and interdependencies. Protecting key elements like SCADA systems, operators and communication networks will no longer suffice as a growing number of new small agents will begin to play active roles in EPSs. These new agents, who will most likely operate based on lower cybersecurity levels, will be easier targets for cyberattacks. Large-scale attacks conducted against thousands or millions of them could lead an

EPS into chaos due to the increasing dependence of EPSs on those small agents. In the long term, those agents should be included in EPS risk analysis and defence strategies.

5. ADDRESSING CYBERSECURITY CHALLENGES

Enhancing EPS cybersecurity requires a broad and diversified range of actions. Grouping them into a few categories, as shown below, can help the analysis.

A. Hard Cybersecurity

Hard cybersecurity refers to mechanisms like access control, authentication, malware control, encryption and other functions commonly used in most computational networks. These are essential cybersecurity tools but are not enough for EPSs: if they fail – and sometimes they do fail – some critical elements of the system may become unprotected. Most hard cybersecurity threats (outdated or poorly configured software, weak passwords, excessive privileges, physical access to critical cybersecurity devices, non-cybersecurity-aware teams, social engineering and many others) are not specific to EPSs and can be fought by well-known strategies. In this work, the hard cybersecurity specificities of EPSs are examined.

One of them is the relative order of importance of the three highest-level goals of cybersecurity [12], namely confidentiality (information is accessed only by authorised agents), integrity (information is changed only by authorised agents) and availability (non-authorised agents cannot substantially harm the behaviour of the network) – the CIA triad. In some business sectors, integrity or confidentiality are often the most important goals. A bank can, in extreme contingencies, temporarily interrupt its online services to avoid interference or damage to its databases; a health insurance company can do the same to preserve the confidentiality of its records. In EPSs, however, availability is paramount and any cyberattack-fighting approach that requires interruption of services is unacceptable.

Cybersecurity policies and strategies must consider the importance of availability and deal properly with associated side-effects. One side-effect is related to the presence of outdated equipment and software co-existing with other equipment in a real-time operational environment. Due to the long lifespan of power system computational hardware and even software, it happens that, during a product lifecycle, suppliers stop providing updates and support or even abandon the market, thus leaving products running outdated versions that are potentially vulnerable to cybersecurity threats. Trying to update these products often brings risks of serious operational problems and raises availability concerns, therefore a common approach is to keep them operating as long as they are performing satisfactorily and to be aware of the risks. To avoid

this uncomfortable and dangerous situation, designers should consider the ease of component replacements from the design phase. Plans to deploy a new component in a system – hardware, software, communication protocols or others – should include a well-documented, simple and smart strategy for its replacement in future.

Another characteristic of EPSs is that they often rely on extensive and poorly monitored communication networks. It is hard to fully prevent physical access to those systems and a single direct connection to a vulnerable point could bypass layers of cyber protection and provide privileged access for attackers. Preventing and monitoring physical access to control and communication hardware is especially important in the presence of outdated hardware or software with insufficient protection against unauthorised accesses, but not only in this case. The possibility of unauthorised direct connections to EPS communications networks should not be neglected and requires appropriate protective measures.

Access to communication networks paves the way for a type of sophisticated attack that has been the subject of much research in recent years: the injection of false data into the measurement network, thus compromising the integrity of the information on which the system's operation is based. This type of attack requires subtle adulterations in some of the field measurements that are received and processed by the state estimator (a software that performs in real time the best possible estimate of the system's state from the measurements received) in order to deceive the supervisory system and take the power system to the state desired by the attacker: unsafe, failure, one that generates undue economic advantages or losses, etc. Detection and prevention of this kind of attack has been the subject of several publications (see [17], for instance).

Other strategies that are not specific to EPSs are especially relevant in this context. Early detection of potentially hazardous behaviour is of great interest and deserves special attention. Honeypots or honeynets developed for real-time control environments can prevent attacks and produce statistics that help refine cybersecurity, and anomaly detection techniques can help to identify suspect behaviours. On the reverse side of the same problem, forensic analysis of attacks (successful or not) is important to retrieve information to improve prevention and to substantiate punitive procedures. Storing enough information for forensic analysis in EPSs, where high rates of information traffic are usual and attacks can take months to prepare, is an issue that merits special attention.

B. Soft Cybersecurity

As seen previously, new "layers" of agents will start to play active roles in EPSs [9]. Many agents, primarily motivated by the expectation of personal profits or advantages and with a significant degree of freedom, will start operating autonomously in power

systems. It can be assumed that some of them will behave in ways that could harm other agents or the whole system, and it is useful to identify them.

In human societies, social mechanisms like reputation and trust reduce the influence of participants that do not behave in a suitable manner; their equivalents in multi-agent systems are the soft cybersecurity mechanisms. The introduction of these mechanisms can be done over solid foundations [13, 14] as they have been used in areas like e-commerce for years.

Reputation evaluation systems, despite some imperfections, have proven effective in motivating agents to behave well and in identifying those that do not. They usually allow parties that have been involved in a transaction to rate each other after its completion. These ratings are then used to construct indexes that are intended to help other agents to decide whether or not to interact with them in future [14].

Differently from reputation, which is built collectively, trust is essentially a personal notion. One agent can even choose to trust another one with a poor reputation, and vice versa. It is also a multifaceted concept that can be split into several classes [13]. The definition of trust that is more appropriate to EPSs is "decision trust" [14, 15]:

> *"Trust is the extent to which one party is willing to depend on something or somebody in a given situation with a feeling of relative security, even though negative consequences are possible."*

This definition indicates that taking risks is an integral part of the concept of trust. It also shows that trust is context-dependent: an agent can choose to trust in another in a specific situation, but not in a different one. The definition also indicates that trust evaluation depends on a balance of potential gains and losses involved in a transaction: when the potential losses are small and the potential gains are significant, one can choose to trust a partner that one would usually not trust. Even if the concept of trust seems to be inherently fuzzy, it needs to be converted to binary values: an agent must decide whether a potential partner is or is not trustworthy enough to engage in a particular transaction.

Application of trust and reputation concepts to the upper layers of future EPSs [9], where thousands or millions of different agents with different capabilities and goals will interact, is not straightforward and is more complex than in the realm of e-commerce. Due to the large number of variables still undefined today, it is debatable whether an early effort to develop attempts at realistic simulations for study purposes would be productive. However, it is likely that the overall system behaviour would benefit from research in this area.

C. High-level Coordinated Actions

High-level coordinated actions are essential components of the defence strategy against some types of cyberattacks, whether directed at EPSs or embracing other critical infrastructure. Inter-agency joint and combined exercises should be centrally coordinated and involve different public and private partners, including those sectors of the armed forces and government responsible for the cyber defence. The engagement of different sectors of society assists in building a robust cybersecurity community capable of exchanging experiences and good practices, and of establishing protocols for information sharing and cooperative work. This may prove essential in crisis situations; otherwise, even the exchange of basic information can be difficult.

There are several initiatives of joint exercises around the world, such as Cyber Europe, a pan-European cyber crisis exercise organised by the European Union Agency for Cybersecurity (ENISA); Cyber Perseu in Portugal, coordinated by the Portuguese Army; the UP Kritis in Germany, conducted by governmental authorities and industry; and the Cyber Guardian in Brazil since 2017, under the coordination of the Cyber Defence Command (ComDCiber). These exercises help to promote collaboration and information exchange at national or supranational levels.

The evolution of EPSs over the next decades and their emerging cybersecurity challenges discussed in previous sections will need to be gradually included in those exercises. They will have an impact on the simulation of scenarios and cyberattacks and bring new vulnerabilities to be reproduced; crisis management, incident response and actions plans must evolve accordingly.

D. Effective Governance in Cybersecurity

Effective cybersecurity governance in EPSs should ideally encompass government, defence, agencies related to EPSs and other critical infrastructure sectors, customers, utilities, private sector representatives, academia and civil society. The digital resilience of EPSs – which is the primary goal of EPS cybersecurity – should be gradually taken to nearly the same level of relevance as EPS energetic supply security or electric operational stability, making cybersecurity a C-suite issue. Cybersecurity managers must also have expertise in topics such as risk and compliance management, corporate governance and overall business objectives. Direct access to senior corporate management is also a must, and all relevant EPS-related agents should adapt to these requirements.

Some important lessons learned indicate risks that should be avoided: (i) excessive securitisation and militarisation of cybersecurity; (ii) exclusion of non-state actors from cybersecurity governance, priority setting and policy-making; (iii) solutions

that seek to block applications, remove content and criminalise behaviours; and (iv) coordination problems within institutions.

The institutionalisation of EPS cybersecurity would ideally encompass technical entities that contribute to the development of related policies, standards and practices. Frameworks for the certification of products, processes and services of interest to EPSs, including concerns with cyber risks brought about by 5G, are also necessary.

Such measures can help to improve the cybersecurity of current EPSs; however, they are insufficient to meet all future needs. The cybersecurity governance of the new layers of autonomous agents is an open issue that deserves special attention as those newcomers will become the weakest link in the cybersecurity chain.

E. Cybersecurity Due Diligence

It may already be a challenge for EPS companies to identify and protect all their critical assets, which can depend on vast, far flung and complex global supply chains. However, the problem is compounded by the ever-increasing degree of digital interconnection with other companies because concerns about cybersecurity can be as different as the companies themselves. For example, a company that builds and operates a set of separate transmission lines (an approach that is part of the EPS business model in some countries) could be much less concerned about cybersecurity than the utilities to which the lines are connected or the national EPS operator. Since the operational networks of those companies are connected to exchange real-time information and commands, a weak link could compromise the cybersecurity of the whole system.

Due diligence of the connection points with other companies is recommended, as well as the definition and enforcement of proper standards to be followed by all parties. And, considering that in some countries the purchase, sale, split and merge of EPS companies are routine, a well-planned cybersecurity due diligence strategy would help provide more agile and orderly evaluation processes.

Extending due diligence strategies to the new layers of EPS agents is a challenge that will probably need to rely on the definition of good and specific connection standards.

F. Staff Awareness and Training (IT and OT)

Sharp differences in cybersecurity approaches do not only occur between different EPS companies; they often exist inside the same utility. Priorities of corporative information technology (IT) staff concerning cybersecurity can greatly differ from those of the teams of real-time operation and SCADA systems (operational technology – OT), and the mutual lack of knowledge about the other environment

brings difficulties and risks. As mentioned before, in real-time control environments the availability of information is much more important than its confidentiality, and even short unplanned interruptions are usually a significant issue that can lead to problems on the energy supply.

The increasing connection between IT and OT environments makes it essential to bridge the gap between their respective cybersecurity teams. In EPS utilities, cybersecurity should be viewed from a broader perspective related to the protection of critical infrastructure, of which the cybersecurity of both operative and corporative networks is part. Drawing up proper awareness and qualification plans for IT/OT professionals should narrow the gap, but it requires a common curriculum that promotes multidisciplinarity. Joint work of IT/OT professionals, as in incident handling teams, is necessary since both environments are increasingly interdependent and connected. Extending this approach to the armed forces or other organs responsible for the cyber defence of critical infrastructure improves their effectiveness because they need well-trained professionals with extensive knowledge of the subjects to be protected who are able to work in cooperation with other experts.

G. Threat Intelligence as a Service

The development of malware, espionage or even cyber weapons is greatly facilitated by the Dark Web [16] and the anonymity that it provides. Effective cyber exploits are monetised and sold in specialised markets, and threat agents that do not have the technical ability required to build specific "tools" can now buy the desired features and hire additional developments.

The development of threat intelligence as a service (ThIaaS), using methodologies such as data mining and machine learning, can help EPS agents to identify, mitigate and prevent attacks, security incidents and other vulnerabilities faster and more efficiently. This service could be leveraged by national or supranational cybersecurity centres and based on an international collaborative environment.

An important support to a network of EPS threat intelligence would be the use of distributed SCADA honeypots and honeynets. Their relevance is expected to increase and, despite the development and monitoring costs involved, they deserve more attention than they have received so far.

H. Research and Development

Research and development (R&D) activities are essential in rapidly evolving technology domains such as cybersecurity. This is even more evident in the case of EPSs, where the physical system itself is changing. Some important research subjects are common to other cybersecurity application domains, like threat intelligence and

topics of artificial intelligence, machine learning, big data analytics and others; other R&D subjects are more specific to current or future needs of EPSs as soft cybersecurity, SCADA honeynets, monitoring of motivations for attackers, visualisation tools for situational awareness etc.

6. CONCLUSION

EPSs are undergoing deep changes driven by forces that can be grouped under the triad of decarbonisation, digitalisation and decentralisation. Some of them are likely to have a strong impact on EPS cybersecurity, such as the multiplication of autonomous agents with active participation in systems and increased vulnerabilities and motivations for attackers.

The new generation of EPS structures is in its infancy, but will hopefully allow the definition and application of satisfactory levels of openness and interoperability with robust cybersecurity in terms of appropriate policies, technologies and processes and well-trained teams. Early actions in this direction could prevent the development of a chaotic and unsafe environment that would resemble the current internet.

In this work, a non-exhaustive list of foreseeable imminent EPS structural changes is presented and discussed from a cybersecurity perspective, including the resulting risks and possible approaches to mitigate them. Accurately predicting all future structural transformations of EPSs and related new cybersecurity challenges and needs is a very difficult task. Nevertheless, developing tools and technologies for effective cybersecurity governance at all layers of new EPS agents and promoting intensive R&D activities to provide technical responses to emerging challenges are some of the right strategic actions to face the huge uncertainties that the industry 4.0 paradigm will bring to EPSs in the near future.

ACKNOWLEDGEMENT

The authors of this paper thank Marcelo Malagutti (PhD Visiting Research Student at King's College London) for his comments, suggestions and revisions.

REFERENCES

[1] M. N. Schmitt, *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*. Cambridge University Press, 2017.

[2] M. Wooldridge and N. Jennings, "Intelligent Agents: Theory and Practice," *The Knowledge Engineering Review*, vol. 10, no. 2, pp. 115-152, 1995.

[3] S. D. J. McArthur, P. C. Taylor, G. W. Ault, J. E. King, D. Athanasiadis, V. D. Alimisis and M. Czaplewski, "The Autonomic Power System Network Operation and Control Beyond Smart Grids," in 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe), pp. 1-7, Berlin, 2012.
[4] K. Geers, "The Cyber Threat to National Critical Infrastructures: Beyond Theory," *Information Security Journal: A Global Perspective*, vol. 18, no. 1, pp. 1-7, 2009.
[5] R. M. Lee, M. J. Assante and T. Conway, "Analysis of the Cyber Attack on the Ukrainian Power Grid," *Electricity Information Sharing and Analysis Center & SANS Industrial Control Systems Report*, March 18, 2016. Available at http://www.nerc.com/pa/CI/ESISAC/Documents/E-ISAC_SANS_Ukraine_DUC_18Mar2016.pdf.
[6] European Commission, "Commission Recommendation (EU) 2019/553 of 3 April 2019 on Cybersecurity in the Energy Sector," *Official Journal of the European Union L 96/50*, 5 April, pp. 50-54, 2019.
[7] S. D. J. McArthur et al., "Multi-Agent Systems for Power Engineering Applications - Part I: Concepts, Approaches, and Technical Challenges," *IEEE Transactions on Power Systems*, vol. 22, no. 4, pp. 1743-1752, 2007.
[8] L. Obregon, "Secure Architecture for Industrial Control Systems," *SANS Institute Information Security Reading Room*, pp. 1-25, 2015. Available at https://www.sans.org/reading-room/whitepapers/ICS/secure-architecture-industrial-control-systems-36327.
[9] G. P. de Azevedo, "Distributed Energy Resources and the Smart Grid: The Role of Cybersecurity," *Accepted for presentation at Cigré 2020 Paris Session*, August 2020.
[10] S. C. da Cruz Junior, "Cyber Security and Defence in Brazil and a Revision of the Strategies of the United States, Russia and India for the Virtual Space (in Portuguese)," Institute of Applied Economic Research (IPEA), Brasília, 2013. Available at http://hdl.handle.net/10419/91261.
[11] M. D. Cavelty, "Critical Information Infrastructure: Vulnerabilities, Threats and Responses," *Disarmament Forum*, UNIDIR, Issue 3, pp. 15-22, 2007.
[12] D. Kapellmann and R. Washburn, "Call to Action: Mobilizing Community Discussion to Improve Information-Sharing About Vulnerabilities in Industrial Control Systems and Critical Infrastructure," In Proc. 11th International Conference on Cyber Conflict: Silent Battle, NATO CCD COE Publications, pp. 1-23, Tallinn, 2019.
[13] J. Sabater and C. Sierra, "Review on Computational Trust and Reputation Models," *Artificial Intelligence Review*, Springer, vol. 24, no. 1, pp. 33-60, 2005.
[14] A. Josang, R. Ismail, and C. Boyd, "A Survey of Trust and Reputation Systems for Online Service Provision," *Decision Support System*, Springer, vol. 43, no. 2, pp. 618-644, 2007.
[15] D. H. McKnight and N. L. Chervany, "The Meanings of Trust," *Technical Report MISRC Working Paper Series 96-04*, University of Minnesota, Management Information Systems Research Center, 1996.
[16] R. Koch, "Hidden in the Shadow: The Dark Web – A Growing Risk for Military Operations," In Proc. 11th International Conference on Cyber Conflict: Silent Battle, NATO CCD COE Publications, pp. 1-24, Tallinn, 2019.
[17] X. Li and K. W. Hedman, "Enhancing Power System Cyber-Security with Systematic Two-Stage Detection Strategy," *IEEE Transactions on Power Systems*, vol. 35, no. 2, 2020.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Towards Classifying Devices on the Internet Using Artificial Intelligence

Artūrs Lavrenovs
NATO CCD COE
Tallinn, Estonia
arturs.lavrenovs@ccdcoe.org

Roman Graf
AIT Austrian Institute of Technology
Vienna, Austria
roman.graf@ait.ac.at

Kimmo Heinäaro
NATO CCD COE
Tallinn, Estonia
kimmo.heinaaro@ccdcoe.org

Abstract: Hundreds of millions of devices are directly reachable by anyone on the Internet. Security researchers and malicious actors are highly interested in ICS, IoT, and building automation and networking devices that can be compromised to negatively affect either a specific person or organization or a whole country at once. The current approach for determining a class of individual device is to conduct a manual investigation or apply static rules to large sets of devices, which is time-consuming and ineffective. We are proposing to utilize neural networks for automated classification.

Many devices have a generic web interface supporting HTTP protocol. We have investigated which features of the HTTP responses from these devices are meaningful for training the neural network model and enabling classification of devices. We have trained neural network models and assessed their accuracy to be 87%. We are analysing the classified sets of the whole Internet scans consisting of tens of millions of devices and comparing them between the years 2018 and 2019 to identify the changes. This kind of all-encompassing view might reveal positive and negative trends that are

happening to specific classes of devices, which might be correlated with real-world events, e.g. new policies issued by governments.

Keywords: *devices on the Internet, classifying devices, machine learning, neural network*

1. INTRODUCTION

Billions of different devices are connected to the Internet and predictions for the next decade expect geometric growth. Statista projects that there will be 75 billion IoT devices by 2025 [1]. The way these devices are connected to networks varies, and only a small portion of all devices on the Internet are publicly reachable by anyone. Unsophisticated actors can access, abuse and exploit reachable devices with known vulnerabilities. Understanding the potential risks and corresponding impacts, or assessing the current state, requires knowledge of classes of devices and their location. Academic and technical research can benefit from this understanding, and it can also provide sufficient background to help policymakers address security concerns regarding these devices.

Identification and classification of reachable devices on the Internet has traditionally been a straightforward process. The targeted protocol port gets tested to check it is open, and possibly a protocol payload is sent and the response processed. Depending on the case, the investigation stops here or continues with additional protocol requests that extract the properties of the devices, possibly identifying the manufacturer or model. If different classes of devices use the targeted port, then classification can be attempted using static rules. Heterogeneity of devices has grown over time, and it has become unfeasible to achieve a high coverage and accuracy rate when classifying large sets of devices. We are attempting to solve this problem by creating a neural network that replaces the static rule stage in the network research.

In Chapter 2, we explore what kind of devices are available on the Internet and why, as well as how they can be classified. Chapter 3 describes our application of machine learning to solve the device classification problem. Chapter 4 explores the results of the classification and compares them between standard and alternative HTTP ports between the years 2018 and 2019. Conclusions and future work are discussed in Chapter 5.

2. DEVICES ON THE INTERNET

In this research, we are attempting to begin to ask what exactly is on the Internet and what the risks are. We are only investigating devices that are reachable on the Internet – reachable meaning that the device receives, processes, and responds to network packets coming from anywhere on the Internet. In general, these packets target specific ports corresponding to known and common protocols. Only a small fraction of all the devices on the Internet are reachable in this way.

A significant number of different services are required to be reachable on the Internet for anyone in order to function properly, e.g., web sites on HTTP and HTTPS, authoritative DNS. Some services are required only for use in a home, office or ISP local network, e.g. DNS resolver, UPnP discovery. The core issue is that the number of devices that are reachable on the Internet far outweighs the number that is required. The leading causes of unnecessary reachable devices are manufacturers' default configurations and network misconfiguration while installing a device.

Reachability significantly increases the attack surface of these devices. Some services can be abused by default, e.g. DNS resolver without rate-limiting for reflected DDoS attacks. Some devices are entirely unprotected while others might contain a publicly known vulnerability that an attacker has to exploit. Depending on the vulnerability, the attacker might achieve a different level of access, from leaking insignificant information up to full control of the device. Depending on the class of the device, the impact of the compromisation can vary drastically. A compromised ICS device can interrupt essential services to vast regions, affecting millions of people, while an unprotected printer might only waste printing toner, causing inconvenience to a single person.

Even if there are protection mechanisms in place like authentication, no immediately abusable services and no known exploitable vulnerabilities, the risk that new vulnerabilities can be discovered in future is ongoing. Unnecessary reachability is already an indication of poor device management practices. No security updates for most devices is the norm; many of the newly installed devices are left untouched until the end of their life for as long as they serve the required purpose.

There are a variety of protocols worthy of investigation for classification. In this research, we are only focusing on the HTTP protocol being utilized on standard port 80 and common alternative port 8080. Implementing a web control panel utilizing HTTP protocol is the cheapest and easiest way that manufacturers can provide a control interface for a device being sold to consumers. This is a ubiquitous protocol supported by every investigated device class, justifying the choice.

A. Classifying Reachable Devices

Multiple approaches suitable for classification of remotely reachable devices exist, but they can all be reduced to acquiring properties of devices and applying a set of static rules to them. The most common property is a check to verify if a specific port or range of ports is open. After this check, port-specific negotiations can occur, and additional information, varying drastically in quality and quantity, can be acquired. At the very least, it can be confirmed if the specific device on the specific port supports the tested protocol. In best-case scenarios, the manufacturer, model, version and even location and purpose of the device can be determined.

After possible properties are acquired and investigated, rules can be developed to match these properties and to locate all matching devices in large data sets, e.g. a full Internet scan. These rules can be something as simple as a unique and rare port being open, up to matching the manufacturer and model returned in the response. These rules are made by humans and usually target common or high impact devices. As many devices require thorough manual investigation to classify them, it is unfeasible that full coverage can be achieved. This is the most common approach for classifying devices in academic and industry research, including device search engines such as Shodan and Censys.

Additional properties can be gathered indirectly by fingerprinting the scanning and communication process or independently by identifying a network, its location and DNS name. These properties are primarily used in the manual investigation of the individual devices and rarely for creating static rules because of the high variability of this data.

This approach has a major drawback. It works perfectly for locating a specific subset of a specific device class using its properties and their values, which are known in advance and were acquired through manual investigation. But what happens when there is a large set of devices or even a single device that has to be classified? The set of available static rules can be applied to it, and there might be a match; in that case, there is no issue. However, if there is no match, then the device is left unclassified and requires manual investigation, which is time-consuming and does not guarantee success. Utilizing a machine learning classifier can solve these types of questions.

B. Related Work

Until recently, classifying devices on the Internet was done in a static way (described in 2.A) both for academic research and industry purposes. Only in recent years have researchers attempted to address this issue using machine learning. Two vantage points are being investigated: reachable device classification using data sets from Internet scanning, and device classification using an observer data set, which includes

all communicating devices, including non-reachable ones. The latter does not provide a full Internet view but provides highly valuable information for internal networks where observer access is possible.

The observer's vantage point enables data to be gathered over long periods of time, from which behavioral profiles can be created. It is also possible to create profiles without decoding the appropriate protocols, and in some cases, it is even impossible because of encryption, e.g., HTTPS. These profiles allow not only the classification of devices but also the identification of misbehaving compromised devices. Sivanathan et al. created a classifier based on existing campus network data that was able to distinguish IoT and non-IoT devices [2]. Bezawada et al. acquired fingerprints from different levels of the same network traffic and combined these into behavioral profiles suitable for machine learning [3].

Yang et al. trained classifiers on data acquired from multitude scanned protocols commonly used by IoT and ICS devices, which were augmented with fingerprints extracted from the network layer communications [4]. This research introduced a significant improvement in labelling training set by the automated scraping of manufacturer and model names of devices from the Internet and matching them against protocol responses in the data set. This developed model has been applied by Jia et al. to determine ownership of devices [5], therefore demonstrating the value of a universal device classifier in helping to solve various research problems.

C. Classes of Devices

Multiple different classifications have been proposed for the devices on the Internet, varying significantly in terms of set size [3], [4], [6]. We propose a small set of 10 classes where every class is selected based on the role, impact, and size of the reachable device set as well as its historical prevalence.

Setting device class definitions is a balancing act, as these can be viewed from the user, functionality, impact and observer perspectives. Creating more classes requires a larger and more precise labelled training set without guaranteed improvement of the total overview. We have identified indistinguishably similar behavior even within small class sets because of the generic HTTP protocol requiring a special class for these devices. At the same time, some of the proposed classes have small subsets of devices, which vary drastically in their behavior and specific purpose. Although the labelled set is significant and proportional to the whole data set, it is not sufficiently representative of various rarer devices and subclasses to train the classifier. When combined with hard-to-distinguish protocol responses, this can introduce even more uncertainty. These issues can be mitigated by augmenting data sets with features from other protocols.

The ICS class contains the most impactful devices which can affect not only individual users but potentially whole regions. It includes industrial control systems, SCADA, and building automation devices. The role and software vary drastically for devices in this class. Although through significant scanning and notification efforts the number of reachable devices has fallen, we are keeping this class.

Network devices are classified as the NET class, which includes all the wired and wireless devices used in individual residential installations and most of the devices serving a more significant role on the network, providing connectivity to organizations and other networks. These are primarily routers, switches, and firewalls. The impact of attacks on these devices cannot be overstated, as not only detectable network interruptions but also hidden MITM attacks can be executed. Other devices in this class include network storage, televisions, and streaming set-top boxes. The INFRA class encompasses data center infrastructure devices affecting the physical properties of the server hardware. These are high-impact devices providing server control panels and virtualization solution control panels.

Although a variety of IoT devices are significant from the serving role viewpoint, we classify all of these in one IOT class. The ratio of IoT devices connected to the Internet versus directly reachable devices is lower than for most other classes. This can be explained by the different ways in which different devices are connected to networks.

The historically prevalent device classes PRINTER, IPCAM, and VOIP are kept separate. These classes had historic public mass attacks that negatively affected a large number of people, e.g. wasting toner printing unwanted documents, leaking private video feeds. Thus their reachability should have decreased over time. The IPCAM class includes not only IP cameras but also DVR and NVR devices that provide recording and viewing functionality. The PRINTER class includes printers and network print servers. The VOIP class includes phone sets, conferencing solutions and VoIP gateways.

It is possible to determine with a high degree of likelihood whether or not a specific device is a generic web server. Features like unsupported HTTP protocol version 1.1, the wrong clock which starts to count time from Unix 0 seconds, and the lack of any headers indicate custom or outdated server software, which usually suggests an embedded device and only in rare cases serves a generic web server role. If we are unable to classify these devices into any other category because response features are insufficient, we classify them as UNCLEAR. This class also includes manufacturers that are represented in multiple classes but where no clear dominant class is established and it is not possible to distinguish device classes from responses, e.g., the same web interface is re-used across classes. In the remaining cases where

we are unable to confirm that the device is not a generic web server, we classify them as UNCATEGORIZED.

From a security research perspective, generic web servers hosting various web applications are often the least exciting class of reachable devices. These devices are much more often properly managed and automatically updated, as they are usually reachable on purpose. The most vulnerable parts of these devices are web applications themselves, not the HTTP servers, but these applications in most cases are reachable using the domain instead of the IP address, which involves a different kind of scanning. There are web applications that are configured to process requests received without the domain name, but quantity-wise they are a minority. We classify all generic web servers, web applications, and services related to these, e.g., CDN, as WEB class.

3. NEURAL NETWORK

The scanning output is HTTP responses that are text in a JSON format. The text classification task in the cybersecurity realm is implemented by a number of text classification methods. Often, classification methods suffer from large vector sizes and are less effective as the number of samples rises. The autoencoder makes use of neural networks which are already in use by latent semantic analysis for text categorization [7] to reduce dimensionality and to improve performance. Another application [8] employs an artificial neural network to improve text classifier scalability. The advantage of the autoencoder method is that it learns automatically from examples.

The main advantage of existing text classification methods, such as Support Vector Machine (SVM) [9], Word Embeddings Neural Networks or the Gensim tool, is that they perform better with a massive database for training to provide meaningful results, and we have a big dataset. However, the common disadvantage of these techniques is the lack of results transparency due to employing vectors containing real-valued numbers. These tools provide results, but it is difficult to explain how the results are calculated. Another disadvantage is the inability to handle unknown words or words which were not included previously in the training vocabulary. The SVM approach is limited by choice of the kernel, which is a general weak point of SVM applications.

Alternative algorithms employing categorical features and labels are Naive Bayes [10], Logistic Regression [11], and Random Forests [12]. Approaches based on decision trees such as Random Forests are very fast to train but quite slow to create predictions once trained. A higher degree of accuracy requires additional trees, which means losing performance. Naive Bayes often serves as a robust method for data classification, but the vectors representing an incident in Naive Bayes are larger than

in word-embedding methods, and also Naive Bayes classifiers make a very strong assumption on the shape of the data distribution. Further problems may result due to data scarcity, which can result in probabilities going towards 0 or 1, leading to numerical instabilities and worse detection results. Logistic regression, like a Naive Bayes method, requires each feature in an incident to be independent of all other features. Logistic regression models are also vulnerable to overconfidence as a result of sampling bias. Consequently, for the particular use case of classifying IoT devices, we suggest using the simplest neural network for text classification that scales well because of the small vector size while maintaining a high level of accuracy.

A. Features Used for Classification

Features of the HTTP responses suitable for the classification have previously been explored by Lavrenovs et al. [13], [14]. For this research, we have decided to use all HTTP response headers and their values, Autonomous system (AS) name, HTML structure hash, body title, body keywords, SSL certificate issuer, and subject.

Specific features are extracted from the response body. HTML tree, in many cases, uniquely identifies groups of the same devices as long as the tree is large enough. To decrease data pollution, we are using only the hash of the HTML tree. The first title is extracted from the HTML body. These titles can often identify specific device models, manufacturers and functionality. The body of the response contains a significant number of mark-up language elements, which do not necessarily benefit us as separate features if the body tree hash is being utilized. We keep only the 1,000 most common words.

Although HTTPS protocol is not being targeted specifically, a small subset of the devices with redirects to HTTPS have numerous TLS properties. However, most of them are usually not uniquely identifying device classes on their own. Even supported ciphers and their order can be used as features, and all of these properties are worthy of investigation in the future for the HTTPS device scan on the Internet. For this research, we use only SSL certificate issuer and subject as those were used for manually labelling the sample and often identified the class of the device on their own.

B. Data Sets

We are operating with four data sets created by scanning the Internet using scanning tools commonly used for research: zmap and zgrab. Both HTTP default port 80 and common alternative port 8080 were scanned in December 2018 and one year apart in December 2019. Up to three redirects are being followed to any port including HTTPS, in which case TLS negotiation is being saved as well. For the standard port in 2018, there are 54,811,827 elements, and in 2019 there are 57,131,825 elements. For the alternative port, there are 7,792,077 and 8,100,201 elements, respectively. An element

is a single response or response redirect chain corresponding to a single request that contains at least one proper HTTP response. Specifically targeting HTTPS ports and also analyzing broken responses would identify additional web control panels, but we have excluded that from the scope of the current research.

We have augmented elements in data sets with additional features. AS name is looked up via the Maxmind GeoIP database. HTML tree hash, first title and body words are all generated from the response HTML body itself.

The labelled set consists of 171,791 elements. It was created from random elements of the 2018 port 80 data set and therefore is unbalanced across classes. There are 132,562 WEB, 22,002 NET, 9561 IPCAM, 711 INFRA, 265 VOIP, 243 ICS, 218 IOT, 153 PRINTER, 4175 UNCLEAR and 1901 UNCATEGORIZED devices in the labelled set.

C. Comparison to the Existing Classification

The overall idea of our solution and [4] is the same: to classify devices on the Internet using artificial intelligence from the remote point of view. The classification model suggested in [4] provides classification on three levels: the type of IoT device, vendor and product. In contrast, the proposed solution aims to classify only by type of IoT device because the vendor and product is just additional information to the class. The approach of crawling additional device information from the Internet, using HTTP queries and analyzing different protocol levels, looks promising but is very unreliable, taking into account the sparse information for such queries. This could be done for the proposed solution as future work, e.g. via query language such as Sparql to compare if this method yields additional value.

Our approach mainly uses information from HTTP headers and body. Yang et al. [4] perform substantial manual pre-training steps. In our approach, we leverage the knowledge and rules developed prior to this research and described in [13], [14]. The existing solution has a very complicated neural network while we propose an alternative solution with possibly more dedicated methods.

Yang et al. classified 15.3 million IoT and ICS devices [4], whereas we analyzed up to 57 million all type devices. Their protocol coverage is higher - 20 protocols (4 ICS). We analyzed HTTP exclusively, but plan to cover additional protocols in the future. Using network-level fingerprinting is extremely unreliable on its own and may produce bias in the overall results. Compared to 41 device types (classes) in the existing research, we make use of 10 classes evaluated from aggregated expert knowledge. The more classes we have, the more unreliable the classification is. The identification of classes itself is a challenging task even for manual analysis and

definition for humans. Therefore, a high number of classes could reduce overall accuracy since there is no common understanding of class definitions.

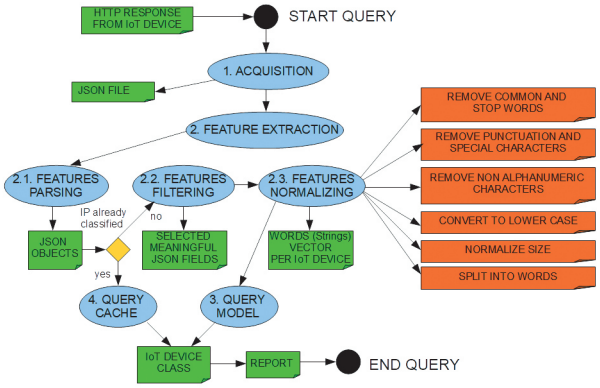
D. General Workflow

Device classification employs features extraction and training of the neural network to produce a model for the queries. Classification predicts previously defined categories for a given sample. There are ten expert-defined classes: ICS, INFRA, IOT, IPCAM, NET, PRINTER, UNCATEGORIZED, UNCLEAR, VOIP, WEB. Supervised learning employs labelled training data to learn mapping functions from a given input (list of words) to the desired output value (class name). A supervised learning algorithm analyzes the data through weights and activation functions that activate neurons and produce an inferred function, which is then used for mapping new samples or correctly determining classification labels for unseen instances.

The workflow process is composed of two parts. One process is neural network model training, where the workflow acquires device data from different sources such as the Internet and domain experts. The model is trained and regularly updated by extended knowledge from new device crawls.

Figure 1 provides an overview of the device classification using neural networks. This approach is based on a knowledge base containing a large number of labelled responses in JSON format (step 1). This data can be provided by different means, collected at different times for particular operating systems, and can be separated by type of application and protocol. The novelty of this approach is that, for typical use cases, we propose to have associated decision rules for initial labelling. All such rules are then aggregated in a common labelled dataset, which supports final classification. We send requests to devices, and the system extracts features (step 2) from the response and stores them for further analysis and queries the model that was trained on the knowledge base. During the feature extraction, we apply parsing, filtering, and normalizing of the content. The final classification result is based on querying the model (step 3) or cache (step 4), if sample hash is already known, and is a report in the form of a particular class name.

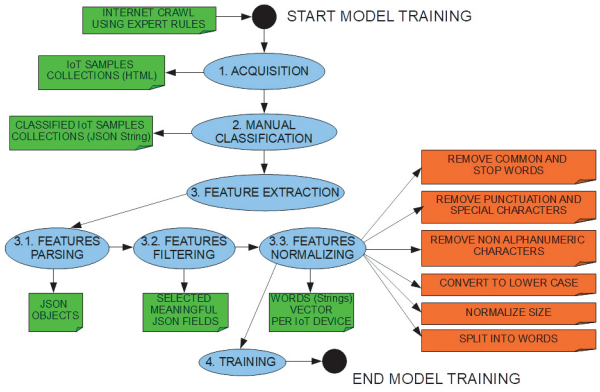
FIGURE 1. THE WORKFLOW FOR FEATURE EXTRACTION AND CLASSIFICATION OF DEVICES USING A NEURAL NETWORK.



E. Model Training

The data for model training is prepared as described in Figure 1 in the previous section. After acquisition and feature extraction, the input for the model is a list of words for each sample. This is then converted into the one-hot vector to be processed in the input level of the neural network model (step 4) in Figure 2. To perform training, features aggregated in text form must be converted into numerical values, since machine learning algorithms and deep learning architectures cannot process plain text. Therefore, each uploaded sample (see Figure 2) is converted into an array of strings, where each string represents a particular feature. Then strings are encoded by indices, and each feature string has a unique index. If this feature repeats in the samples, we re-use its index. Finally, arrays of indexes are converted in one-hot encoded vectors, meaning that the position of each feature in the original feature set is encoded using "1" if a feature exists in the given place or "0" if not.

FIGURE 2. THE WORKFLOW FOR MODEL TRAINING FOR DEVICES USING A NEURAL NETWORK APPROACH.



The neural network used a total of 517,642 parameters during the training. A summary of the neural network training is presented in Table 1. The neural network is composed of an input layer and an output layer. The number of neurons in these layers ranges from 10 to 512. The input layer uses a rectified linear unit (ReLU) as an activation function. The output layer employs a softmax activation function, which provides probabilities as to which of 10 classes a particular sample belongs to.

TABLE 1: SUMMARY OF THE NEURAL NETWORK TRAINING PROCESS.

Layer	Type	Activation Function	Neurons #	Parameters #
Input layer	Dense	ReLU	512	512512
Output layer	Dense	Softmax	10	5,130

We performed a total of 10 training iterations (epochs). The neural network training and accuracy calculation process took 15.723163 seconds (Figure 3). This figure shows that loss and validation loss decreased and accuracy and validation accuracy increased with each epoch.

FIGURE 3. ACCURACY AND LOSS CHARACTERISTICS BY NEURAL NETWORK TRAINING.

We trained two models - one with the full labelled data set (large) and one balanced model (small). Comparing their accuracy (about 87% for small and 97% for the large data set), we noticed by randomly sampling the classified output of the whole data set that the small model performed better due to the bias in the large data set. As the full labelled data set primarily consists of WEB devices, the classified output is significantly skewed towards classifying devices as WEB. To avoid bias

of overrepresented classes in the labelled data set (in total 171,791), such as WEB, we employ a balanced labelled training set (in total 11,479): ICS:243, INFRA:711, IOT:218, IPCAM:1,999, NET:2,000, PRINTER:153, UNCATEGORIZED:1,901, UNCLEAR:1,999, VOIP:265, WEB:1,999. The labelled training data set was divided into a training set (5,628), validation set (2,413), and test set (3,447). The test accuracy is 0.87277.

4. RESULTS

The model was trained using the 2018 standard port labelled data set and applied to the 2019 standard port data set as well as the port 8080 data sets for both years. Although the reachability of devices has been recognized as a poor and high-risk management practice, there was an increase in the data set sizes in 2019.

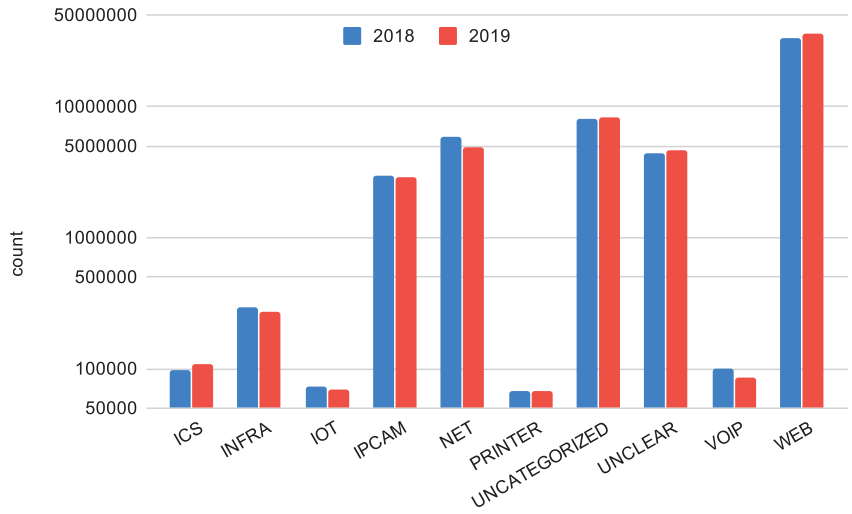
The standard port 80 classification results are provided in Figure 4. As expected from the labelled set, WEB devices were the most prevalent ones. It was not expected that the UNCLEAR and UNCATEGORIZED devices would be so numerous, but that can be explained. UNCLEAR and UNCATEGORIZED devices often have a small set of rare features extracted from the HTTP responses, which makes even manually classifying them challenging and in many cases impossible. While creating the labelled set, many of these devices were categorized. This was done through numerous weak rules utilizing only the available features. These features might be sufficiently rare and unique to not be applicable to the whole data set, in which case HTTP response data on its own might not suffice for accurate classification.

We can observe a slight decrease in reachable INFRA and IOT devices in 2019. As the number of IOT devices is growing significantly, it would be expected that the number of reachable devices would grow over the one-year period. However, this class of devices is the only one of the defined classes that historically could rarely be connected in a way that made them reachable. A more significant decrease in VOIP could be explained by changes in the way this type of device is deployed and managed at the vendor level.

From the publicly well-known attacks targeting IPCAM and PRINTER devices, it could be expected that the number of reachable ones would decrease significantly, but no such trend is observable. One explanation is that the number of newly added reachable devices closely matches the ones that were mitigated. It is currently not clear what portion of these almost 3 million IPCAM devices have to be reachable for remote surveillance and recording purposes.

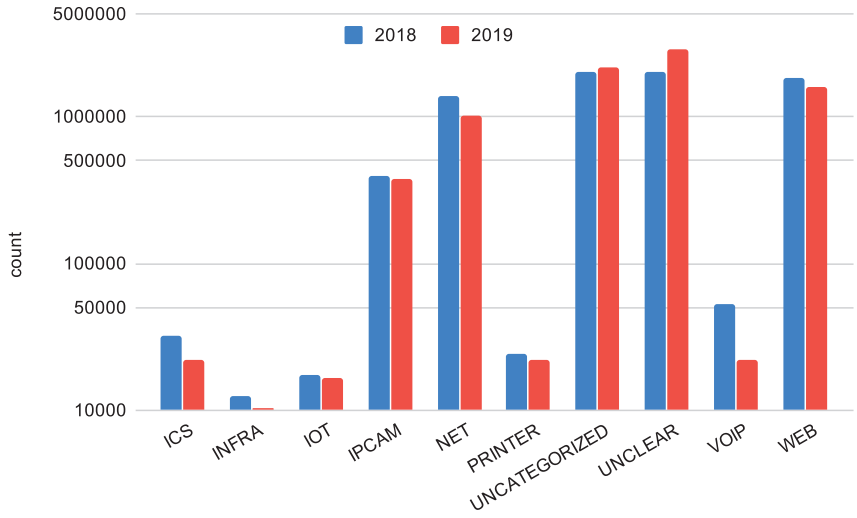
A large number of NET devices was expected. A residential Internet connection device can expose the control panel to the Internet even if the initial setup is done by the ISP technician. A significant drop in the number of these devices might suggest that the device life cycle could be playing a role, with older ones getting replaced and newer ones having a better configuration.

FIGURE 4. DISTRIBUTION OF DEVICE CLASSES FOR PORT 80 FOR 2018 AND 2019.



The alternative port 8080 classification results are presented in Figure 5. As expected, the WEB devices are a proportionally smaller class than on the port 80 where generic websites usually reside. UNCLEAR and UNCATEGORIZED are the two largest classes and show significant growth over the one-year period, which might suggest that the feature difference is significant enough between the two ports that the model needs to be augmented with the alternative port data as well. We can observe much more significant proportion changes among the classes on the alternative port.

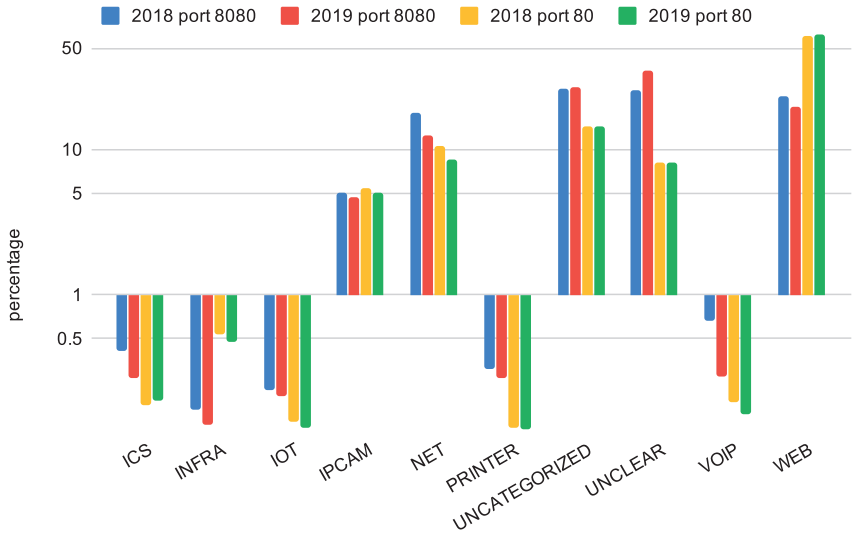
FIGURE 5. DISTRIBUTION OF DEVICE CLASSES FOR PORT 8080 IN 2018 AND 2019.



The relative class distribution for all four classified data sets is presented in Figure 6. This view enables us to make a comparison between the utilization of different devices on different ports. There are other discernible differences besides the already identified WEB, UNCATEGORIZED and UNCLEAR classes. INFRA devices are proportionally about four times less prevalent on the alternative port; this could be explained by the fact that there are a small number of manufacturers whose devices were identified and labelled on the port 80. These devices might use the default port setting, and there might be unidentified INFRA devices defaulting to 8080 port.

Interestingly, IPCAM has almost the same proportion across the ports with the same decrease over the one year. Proportionally, there are significantly more PRINTER devices on the alternative port, and that is explainable with the high variance of device models and default configurations even among individual manufacturers. VOIP, ICS, IOT and NET devices are also proportionally more represented on the alternative port. This might be the result of manufacturers' concerns about creating port conflicts on a single device. This concern is especially valid for NET devices, which are handling networking traffic and possibly forwarding the port 80 to another device.

FIGURE 6. PROPORTIONAL DISTRIBUTION OF DEVICES FOR PORT 80 AND 8080 IN 2018 AND 2019.



5. CONCLUSIONS

We have successfully trained a machine learning classifier for web interfaces achieving 87% test accuracy without the use of a rule engine. Although using the full labelled set to train the neural network achieved higher test accuracy of 97%, further research is needed to determine if this higher accuracy can be achieved while avoiding the bias caused by an unbalanced data set. A large proportion of devices being classified as UNCLEAR and UNCATEGORIZED was unexpected but explainable and can be addressed through augmenting data with features from other protocols. Although the model for the standard port functioned for the alternative port, the increase in UNCLEAR and UNCATEGORIZED devices indicates that there might be a sufficient number of devices unique to the alternative port. This therefore requires the data from the alternative port to be included into the labelled training set or a separate model created.

Our future work will include augmenting the model with HTTPS web interfaces and additional common or high impact port checks and appropriate protocol communication responses. Reverse and forward DNS as an additional source of features could more precisely filter out WEB servers that are currently UNCATEGORIZED. Fingerprinting TCP communications as an additional feature is worthy of investigation as well. Redeveloping rules used for labelling the sample set into a rule engine should significantly increase the accuracy of the classification.

This type of classifier could provide the full Internet view of the reachable devices, with details of individual countries and networks. It has significant value not only for research purposes but also to provide overview reports to decision-makers about which security concerns require the most attention. The same classifier can also be used for internal networks, by-passing firewall restrictions and classifying devices with open ports, thus competing with the observer approach.

The application of machine learning to various research problems is currently hard to replicate in most cases. We are planning to develop the classifier with the discussed improvements as an API available to researchers to help others to address a vast range of network-related research questions more precisely.

REFERENCES

[1] Statista Inc., "Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025," November 2019. Available: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/ [Accessed: 16-12-2019].

[2] Sivanathan, A. *et al*., "Characterizing and classifying IoT traffic in smart cities and campuses," in *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Atlanta, GA, May 2017, pp. 559–564, doi: 10.1109/INFCOMW.2017.8116438.

[3] Bezawada, B., M. Bachani, J. Peterson, H. Shirazi, I. Ray, and I. Ray, "Behavioral Fingerprinting of IoT Devices," in *Proceedings of the 2018 Workshop on Attacks and Solutions in Hardware Security - ASHES '18*, Toronto, Canada, 2018, pp. 41–50, doi: 10.1145/3266444.3266452.

[4] Yang, K., Q. Li, and L. Sun, "Towards automatic fingerprinting of IoT devices in the cyberspace," *Computer Networks*, vol. 148, pp. 318–327, Jan. 2019, doi: 10.1016/j.comnet.2018.11.013.

[5] Jia, Y., B. Han, Q. Li, H. Li, and L. Sun, "Who owns Internet of Things devices?," *International Journal of Distributed Sensor Networks*, vol. 14, no. 11, p. 155014771881109, Nov. 2018, doi: 10.1177/1550147718811099.

[6] Cvitić, I., D. Peraković, M. Periša, and M. Botica, "Novel approach for detection of IoT generated DDoS traffic," *Wireless Networks*, Jun. 2019, doi: 10.1007/s11276-019-02043-1.

[7] Yu, B., Z. Xu, and C. Li, "Latent semantic analysis for text categorization using neural network," *Knowledge-Based Systems*, vol. 21, no. 8, pp. 900–904, Dec. 2008, doi: 10.1016/j.knosys.2008.03.045.

[8] Lam, S. L. Y. and Dik Lun Lee, "Feature reduction for neural network-based text categorization," in *Proceedings. 6th International Conference on Advanced Systems for Advanced Applications*, Hsinchu, Taiwan, 1999, pp. 195–202, doi: 10.1109/DASFAA.1999.765752.

[9] Auria, L. and R. A. Moro, "Support Vector Machines (SVM) as a Technique for Solvency Analysis," *SSRN Journal*, 2008, doi: 10.2139/ssrn.1424949.

[10] Manning, C. D., P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York: Cambridge University Press, 2008.

[11] Cox, D. R., "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society: Series B*, vol. 20, no. 2, pp. 215–242, 1958.

[12] Tin Kam Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, Que., Canada, 1995, vol. 1, pp. 278–282, doi: 10.1109/ICDAR.1995.598994.

[13] Lavrenovs, A. and G. Visky, "Exploring features of HTTP responses for the classification of devices on the Internet," presented at the 2019 27th Telecommunications Forum (TELFOR), Belgrade, Serbia, Nov. 2019, doi: https://doi.org/10.1109/TELFOR48224.2019.8971100.

[14] Lavrenovs, A. and G. Visky, "Investigating HTTP response headers for the classification of devices on the Internet," presented at the 2019 IEEE 7th IEEE Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE), Liepaja, Latvia, Nov. 2019, doi: 10.1109/AIEEE48629.2019.8977115.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Hacking the AI - the Next Generation of Hijacked Systems

Kim Hartmann
Conflict Studies Research Centre
kim.hartmann@conflictstudies.org.uk

Christoph Steup
Anhalt University of Applied Sciences
christoph.steup@hs-anhalt.de

Abstract: Within the next decade, the need for automation, intelligent data handling and pre-processing is expected to increase in order to cope with the vast amount of information generated by a heavily connected and digitalised world. Over the past decades, modern computer networks, infrastructures and digital devices have grown in both complexity and interconnectivity. Cyber security personnel protecting these assets have been confronted with increasing attack surfaces and advancing attack patterns. In order to manage this, cyber defence methods began to rely on automation and (artificial) intelligence supporting the work of humans. However, machine learning (ML) and artificial intelligence (AI) supported methods have not only been integrated in network monitoring and endpoint security products but are almost omnipresent in any application involving constant monitoring, complex or large volumes of data. Intelligent IDS, automated cyber defence, network monitoring and surveillance as well as secure software development and orchestration are all examples of assets that are reliant on ML and automation. These applications are of considerable interest to malicious actors due to their importance to society. Furthermore, ML and AI methods are also used in audio-visual systems utilised by digital assistants, autonomous vehicles, face-recognition applications and many others. Successful attack vectors targeting the AI of audio-visual systems have already been reported. These attacks range from requiring little technical knowledge to complex attacks hijacking the underlying AI.

With the increasing dependence of society on ML and AI, we must prepare for the next generation of cyber attacks being directed against these areas. Attacking a system through its learning and automation methods allows attackers to severely damage the system, while at the same time allowing them to operate covertly. The combination

of being inherently hidden through the manipulation made, its devastating impact and the wide unawareness of AI and ML vulnerabilities make attack vectors against AI and ML highly favourable for malicious operators. Furthermore, AI systems tend to be difficult to analyse post-incident as well as to monitor during operations. Discriminating a compromised from an uncompromised AI in real-time is still considered difficult.

In this paper, we report on the state of the art of attack patterns directed against AI and ML methods. We derive and discuss the attack surface of prominent learning mechanisms utilised in AI systems. We conclude with an analysis of the implications of AI and ML attacks for the next decade of cyber conflicts as well as mitigations strategies and their limitations.

Keywords: *AI hijacking, artificial intelligence, machine learning, cyber attack, cyber security*

1. INTRODUCTION

Artificial intelligence (AI) has been applied in many scenarios in recent years, and this technology is expected to establish itself in further fields over the next decade. Within the military sphere alone, AI technology is expected to penetrate into areas such as intelligence, surveillance, reconnaissance, logistics, cyberspace operations, information operations (the most prominent technology is currently "deepfakes"), command and control, semiautonomous and autonomous vehicles and autonomous weapon systems. Numerous reports and analyses suggest that an AI arms race has indeed already begun [1]. In addition to the military application scenarios, AI systems are also utilised in applications such as public security surveillance [2], financial markets [3], healthcare [4], Human-Computer and Human-Machine Interactions, cybersecurity, power grid management [5], autonomous driving and driver assistance systems. Any of the aforementioned application scenarios are of high value to civilian, governmental or military units and have a high significance to society. Therefore, these applications and the systems involved must be considered as highly valuable assets in cyberwarfare and protected accordingly.

The security of AI systems is currently underrepresented in public discussions; however, reports on successful attacks on AI systems have emerged over the past couple of years. The utilised attack vectors range from requiring little technical

expertise to attacks involving detailed knowledge of the underlying AI [6]. Reported results have ranged from the AI mistaking a turtle for a rifle, to making individuals undetectable to the system.

The penetration of AI throughout digital spaces is likely to increase even further over the next decade, as well as our reliance on its correct identification and reasoning abilities. AI is envisioned to outperform humans in most tasks involving processing large amounts of data/information, high precision or complex reasoning. It is assumed to deliver unbiased and rational results without interference from non-logical events or circumstances. This presumption renders hijacked AI systems an extremely dangerous threat to modern societies.

The wide-range of applications involving AI is startling, especially as AI has been regarded as being almost impossible to secure [7]. In December 2019, Microsoft published a series of materials on the topic, stating that "[i]n short, there is no common terminology today to discuss security threats to these systems and methods to mitigate them, and we hope these new materials will provide baseline language […]" [8]. Over the past decade, we have witnessed increasing and incautious utilisation of AI and ML techniques in applications whose correct functioning is crucial to modern societies. It is easy to imagine how any malfunctioning of these systems might have a devastating impact on civilian lives, financial markets, national security and even military operations.

With society's increasing dependence on ML and AI, we must prepare for the next generation of cyber attacks being directed against these systems. Attacking the system through its learning and automation methods allows the attackers to severely damage the system by altering its learning outcome, decision making, identification or final output. Furthermore, it is difficult to analyse AI systems post-incident and integrate real-time monitoring during their operation: much of the learning and reasoning is done in what is called a "hidden layer" and in its essence corresponding to a black box model. Therefore, the discrimination of a compromised from an uncompromised AI system in real-time is still considered very difficult. With its increasing utilisation in crucial application scenarios, the security of AI systems becomes indispensable.

Knowledge of AI systems' vulnerabilities may also become of high importance to defensive cyber operations. During 2019, we witnessed increasing weaponisation of AI, often to create "deepfakes" – artificially generated or altered media material found to impose a sincere threat to democracies [9]. The uprising of deepfakes has encouraged the U.S. DARPA to spend $68 million on the identification of deepfakes over the past four years [10]. While it is of utmost importance to identify AI-supported disinformation campaigns, identification alone will not stop such operations. Offensive

technological knowledge of how to stop AI-supported attacks will become essential to establish and uphold cyber power in an ongoing AI arms race.

The aim of this paper is to foster understanding of the susceptibility of AI systems to cyber attacks, how incautious utilisation of AI and ML may make societies vulnerable, and to transfer the value of knowing AI-/ML-system vulnerabilities within the ongoing AI arms race. Attack surface modelling is a key contribution to assessing a target's susceptibility to attacks. However, AI systems have several peculiarities, which must be addressed when deriving the attack surface. Within this article, attack surfaces of different AI systems are derived that consider systems' data assets, processing units and known attack vectors, allowing us to understand these systems' vulnerabilities. Furthermore, these attack surfaces must be discussed with the systems' societal and economic impact in mind to allow strategic and policy recommendations. At the time of writing, neither the AI systems' concrete attack surface definition nor the embedment of the different AI systems' specific operational setup have been part of the security assessment of these systems. Allowing an AI-specific, concrete attack surface discussion, which includes the operational setup associated with the AI/ML method utilised by the system, is the main contribution of this article in addition to providing insights into the role of AI systems' susceptibilities to cyber attacks in the next decade of cyber conflicts.

This paper will continue as follows: we start by giving a brief introduction to selected AI and ML methods currently deployed (section 2). We report on state of the art attack patterns directed against these systems and how it must be expected that these systems will become prominent targets over the next decade. We derive and discuss how attack surfaces may be modelled for AI systems (section 3). In section 4, we apply the previously derived attack surface model to AI systems utilising the different methods previously introduced in section 2 to compare their susceptibility to attacks. We conclude with an analysis of the implications of AI and ML attacks for the next generation of cyber conflicts and recent mitigation strategy attempts (section 5).

2. AI AND ML METHODS

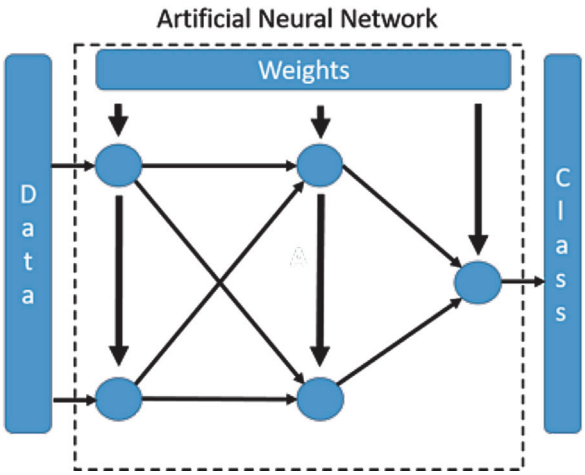
The field of artificial intelligence and especially the sub-field of machine learning is vast. Within the scope of this article, we consider some of the prominently utilised methods with cross-domain applications. Artificial Neural Networks (ANNs) describe the basic principles of neural networks and are commonly applied to predictive modelling problems involving the analysis and classification of non-linear relationships within datasets. Convolutional Neural Networks (CNNs) are an adaptation of ANNs specifically designed to map image data to an output class.

CNNs are commonly applied in prediction problems involving data analyses. GANs (Generative Adversarial Neural Networks) have become publicly renowned through the emergence of "deepfakes", which has yielded strong interest in deep learning methods. Opposing to the discriminative learning of ANNs and CNNs having a clear goal, generative modelling helps with understanding data and generating hypotheses. Support Vector Machines (SVM) were the standard solution to pattern recognition tasks prior to the emergence of neural networks and were used extensively in audio, video and handwriting recognition tasks.

In the next subsections, each of these will be explained briefly to allow for better understanding of security analysis of systems utilising these methods.

A. Artificial Neural Networks

FIGURE 1. EXAMPLARY ARTIFICAL NEURAL NETWORK (ANN). This network consists of three layers with a maximum width of the layers of two (corresponds to the amount of neurons in a single layer). The dots represent the neurons. The arrows from left to right indicate the data flow from the input on the left to the output on the right. The arrows from the top indicate the configuration of each neuron with weights, which were typically acquired using a training phase. The weight collection reflects the learning outcome.



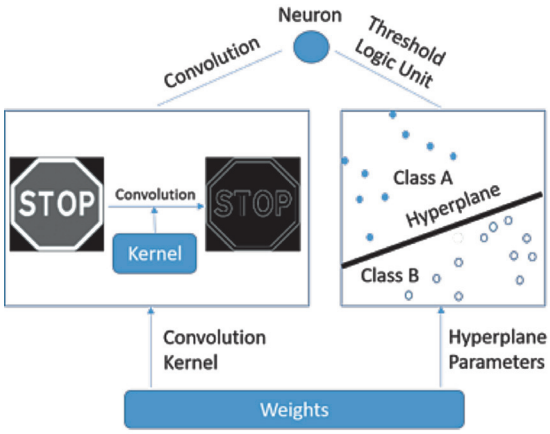
ANNs provide an abstract replication of the processes existing in the human brain. These models consist of simple atomic components called neurons, which are very limited in their individual capabilities, but which may be combined to perform more complex tasks. ANNs usually do not incorporate any task-specific rules, but instead derive the correct output from examples. Similarly to the biological model that inspired ANNs, a simple neuron may only be able to decide if an input is above a certain

threshold or not. However, collectively, a circuit of multiple neurons is capable of performing much more complex tasks. As an example, given a set of panda pictures, the ANN is able to extract a pattern of these pandas. It learns the characteristics extracted from the examples given. The system utilising the ANN will then be able to evaluate any picture with regard to these characteristics, resulting in a "match" if a sufficient number of the characteristics are met and a "mismatch" otherwise. This is called classification. Some systems are also able to provide a confidence ratio for a performed classification. However, the correctness of the classification depends greatly on the amount and variance of the training data provided. In the above example, if the panda training set only contained pandas shown from behind, the system would not be certain of the correct classification of a panda shown from the front, or may mistake an advertising pillar with a poster of black and white dots for a panda.

The peculiar strengths of ANNs are scalability and flexibility, achieved through the combination of multiple neurons. The computational capabilities are achieved through the vast connections between individual neurons. However, these multiple neurons artificially expand the "parameter space" – the space of all possible parameter combinations. Hence, the enhanced flexibility and scalability come at the price of larger training sets and higher computational power being necessary to make the neural network converge towards the correct solution.

B. Convolutional Neural Networks

FIGURE 2. CNN INVOLVING A CONVOLUTIONAL AND DENSE LAYER. The left side shows the operations of the convolutional layers, which perform the data pre-processing and feature extraction through convolution. The right side depicts the dense layers' operations that enable the CNN to classify the data based on the previously extracted features. In this example, a hyperplane is used for the classification.

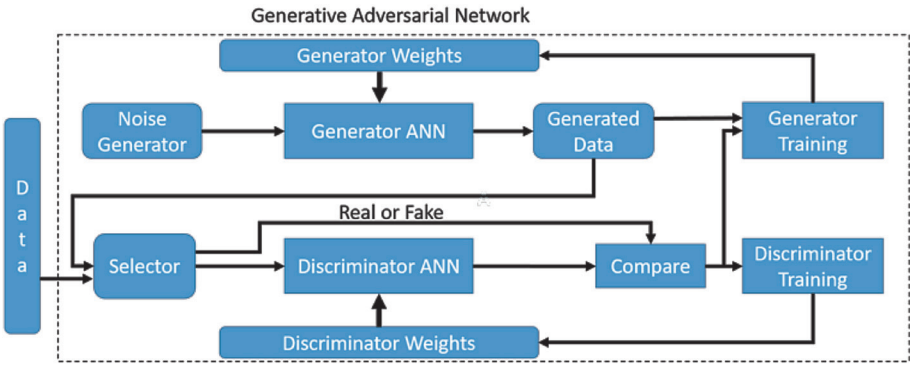


Convolutional Neural Networks (CNNs) belong to the class of "deep neural networks" (DNNs). DNNs are ANNs with multiple layers between the input and output layers. CNNs utilise two types of layers: convolutional layers and dense layers. Within the convolutional layers, each neuron processes only a small region of the input image. However, the regions are partially overlapping. This enables the network to exploit hierarchical patterns within the data and allows it to perform pre-processing and feature extractions. The dense layers are usually fully-connected ANNs used to identify patterns in the output of the convolutional layers. Dense layers are very powerful and induce a large parameter space due to the large amount of weights induced by the inter-neuron connections.

Although the convolutional layers reduce the overall parameter space, typical object detection (image classification and localisation) CNNs, such as YOLO [11], still contain over 60 million parameters. Due to the size of the parameter space, comprehensive training datasets and computational power are needed to train the network sufficiently. Therefore, pre-trained networks are available that may be used and where only the final layers must be modified to adapt to an application specific classification. This process of using pre-trained models is called "transfer learning" and is widely used.

C. Generative Adversarial Networks

FIGURE 3. VISUALISATION OF A GAN. Internally, a GAN consists of two ANNs, the generator and the discriminator, which are trained within a competitive, internal process. The generative network synthesises artificial data from random input, while the discriminator attempts to distinguish real data from the synthetic data of the generator. The selector arbitrarily selects either real or generated data and forwards this to the discriminator. The result of the discriminator is evaluated against the truth given by the selector - the evaluations outcome is utilised to train the generator and discriminator. As a result, two ANNs are trained in parallel: one produces data similar to the training data while the other is capable of identifying synthesised data.

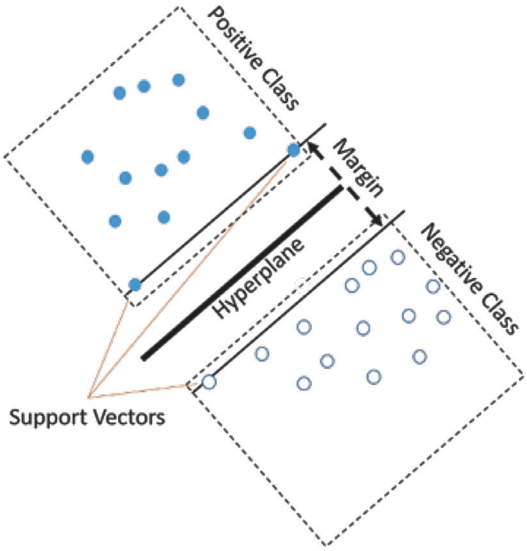


Generative Adversarial Networks (GANs) have gained much attention during the last year due to their frequent utilisation in the creation of "deepfakes". GANs consist of two competitive internal ANNs – the generator and the discriminator. These ANNs are trained in parallel in a competitive manner, which is often deployed as a zero sum or adversarial game. The discriminator tries to detect whether an input is originating from a training dataset or has been synthesised, while the generator generates adversarial samples to mislead the discriminator.

As the competitive training automatically generates feedback information, GANs do not necessarily need labelled training data. However, in order to provide reasonable output, at least the discriminator should be pre-trained on labelled data. For the creation of deepfakes, conditional GANs (cGANs) are often used, which rely on labelled data to allow a target-oriented training.

D. Support Vector Machines

FIGURE 4. VISUALISATION OF AN EXAMPLE SVM. The SVM seperates two classes of data points (blue and white) through a hyperplane while maximising the margin between the hyperplane and the nearest data points. These data points are called support vectors.



Support Vector Machines (SVMs) utilise labelled data and machine learning algorithms to perform classification and regression analysis with the help of a separating hyperplane and cluster support vectors (see Figure 4). SVMs have played a

dominant role in AI systems prior to the rise of ANNs, especially in the fields of text classification and speech recognition.

SVMs utilise mathematical concepts to define a separating hyperplane for a given set of data. Finding a separating hyperplane for a set of linearly separable clusters can be achieved through logistic regression. In order to understand non-linear relationships or solve higher-dimensional tasks, SVMs utilise "kernel tricks". The results achieved by SVMs are considered to be trustworthy and robust. However, SVMs can only perform two-class classifications (i.e. the data can only be distinguished into two categories). If more than two classes exist, algorithms must be applied that reduce the multi-class problem to several two-class problems and SVMs must be trained and executed in parallel. This limitation originates from the definition of a hyperplane, which is utilised to separate two distinct clusters. However, choosing the hyperplane to have a maximum distance between itself and the data clusters yields an inherent robustness against noise.

Some of the drawbacks of SVMs are the limitation to two-class-problems, the complexity associated with reducing multi-class problems to concurrently executable two-class-problems, the utilisation of rather complex mathematical models of kernel-functions, the necessity of labelled data input and difficulties associated with the model parameter interpretation (amongst others: finding the actual kernel function). However, SVMs are still used in various application scenarios stemming from the fields of data science, data analytics and business analytics.

3. ATTACK SURFACE

The security of AI systems and attacks directed against these systems are currently being neglected in public discussion, while the versatile utilisation of AI in varying application contexts is widely discussed. However, within the academic and technical communities, several techniques and attack vectors directed against AI systems and methods have been reported.

Currently, the most prominent attack vector categories are [12]:

- Adversarial inputs;
- Data poisoning attacks;
- Model stealing techniques.

Further attack vectors that have been identified are: model poisoning [13], model and data theft [14], data leakage [15] and neural network trojans [16]. Attack vectors

directed against the AI systems' deployment or training environment are equally applicable. These may be attack vectors directed against servers, databases, protocols or libraries utilised within the AI system. In order to allow a discussion of the vulnerabilities of AI systems, a common understanding of its attack surface must be achieved.

An attack surface allows analysts to depict the means by which an attacker may enter, extract data or manipulate the system in question. It is usually performed on software components, applications or networks in order to understand, assess and manage security risks during the design and development phase. Attack surfaces are usually designed to depict threats to a specific component or application (i.e. ignoring operators or system security issues) that stem from an outsider. However, the concept is also applicable to evaluate exposure to internal attacks [17]. Knowledge of the attack surface is invaluable in order to understand the correlations between exposure, risk and vulnerabilities [18].

A recent report of the Transatlantic Cyber Forum provided a generic, abstract attack surface claiming to cover any ML methods [19]. Opposing the attack surface derived in the aforementioned report, we will follow the OWASP guidelines on attack surface modelling, which yields an abstract yet more concrete attack surface to specific AI systems.

Currently, AI systems often lack sufficient security evaluations [20]. This may be a result of the mutually independent development of AI methods and their implementation in applications: while the application should have a security evaluation, the incorporated AI (utilised by the application through APIs or frameworks) is rarely considered in terms of its security vulnerabilities by the application developers. While the AI framework developers may follow coding standards and guidelines for secure software development, they will not evaluate the potential attack surface of an AI system utilising the framework.

As AI is expected to become ubiquitous over the next decade, the importance of understanding the vulnerabilities of AI systems and methods becomes clear. Within the following subsections, we define how attack surface modelling for AI systems should be done to include the peculiarities of these systems.

A. Data Assets

The attack surface provides information of possible entry points for an attacker as well as exit points allowing access to the systems' data. It is the result of all possible attack vectors against a system or component.

AI systems are data-driven systems that strongly depend on the data quality, authenticity and availability. Hence, data security is of particular relevance when assessing the attack surface of an AI system. Data security is usually evaluated by assessing the input validation, security at rest and security in transition. Assessing these three involves an evaluation of the impact of an attack and its likelihood of occurring. Several attack vectors directed against specific data assets in AI systems have been described (see according subsection of section 4). In addition to the AI/ML specific attack vectors, there have been reports of attacks directed against the databases holding the data assets, yielding data disclosures [21].

The impact of data alterations depends on the AI and ML methods used. Reports of minor alterations yielding majorly false classification with enormous effects in AI systems have been reported [22], while at the same time, some systems are almost ignorant to changes. Overall, the usage of sparse datasets renders the AI prone to adversarial attacks after training [23].

Furthermore, it must be recalled that for modern applications, the AI system is likely to be developed to enable concurrent processing – especially when processing large or complex data, as is the case in most AI application scenarios. A concurrent operation on data assets, however, implies the necessity for data management. The concurrent operation may either be achieved through shared databases or distributed data.

Using a database requires separate securing of that database, especially when utilising distributed and parallel computing, as the database will be addressable (through the TCP/IP stack) for external requests.

Allowing distributed data implies that the data must be kept consistent throughout the system processing entities. This is usually done by a periodic or event-triggered merging of the distributed data assets, where the data is collected from all entities. This requires authenticity of the entities involved and methods to ensure that no manipulation of the data can be performed during transportation (man-in-the-middle attacks).

B. Processing Units

Processing units within AI systems are units that are directly involved in the learning process, the data gathering or the decision making. While some attacks against the processing units will utilise data to perform the attack, other attack vectors may deploy techniques directed against the application involved (e.g. a web crawler used for data gathering is susceptible to web application vulnerabilities), the process itself or the libraries used.

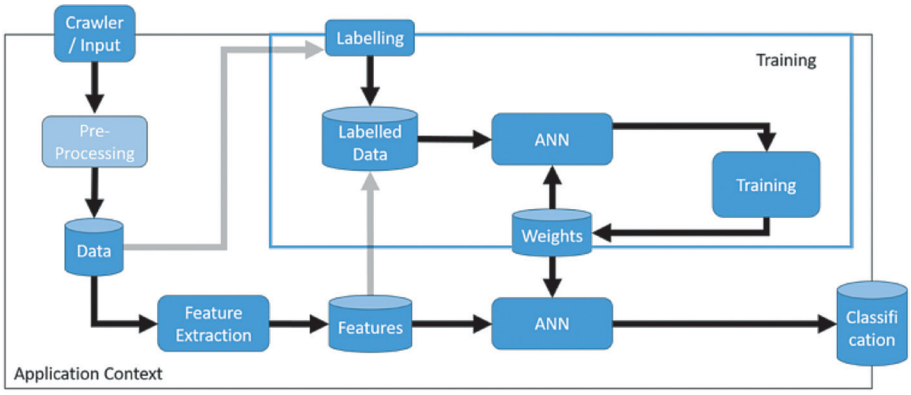
A specific type of attack combines the use of poisoned data and known vulnerabilities in the processing entities [24]. Previous attacks of this type have used audio/video files to hide malicious background operations in a steganographic manner to allow for the execution of arbitrary code. While initially considered as an attack against a specific media player, this attack utilised a meta language library vulnerability. This attack vector could have affected other applications calling the library equally, such as AI systems processing a manipulated file.

4. AI SYSTEM VULNERABILITIES

Within this section, we will use the attack surface considerations made in section 3 to define the attack surface of AI systems deploying the AI and ML methods discussed in section 2. Following the OWASP guidelines on attack surface assessments, we identify entry and exit points and briefly discuss reported and plausible attack vectors. As there are some similarities regarding the attack surfaces of ANNs, CNNs and GANs, a full explanation of an identified attack vector is given at its first encounter only. The summarising conclusion of the findings below is embedded in the overall conclusion and outlook of this paper and given in section 5.

A. ANNs

FIGURE 5. A CLASSIFICATION APPLICATION UTILISING A GENERIC ANN. The incoming data is preprocessed (reduce noise/selection of relevant material) and features are extracted. The data is labelled manually or automatically during the preprocessing. The weights of the network are adapted during the training. The final classification uses the weights derived during the training.



Looking at the overview given in Figure 5, the following attack surface points and associated vulnerabilities are identifiable:

- Crawler/Input – Entry point
  Risk of introducing unscrutinised data, data corruption and poisoning attacks. Crawlers working in a web context are web applications and susceptible to common web application vulnerabilities [25].
- Labelling – Entry point, two cases to be considered:
  Manual annotation: consideration of annotation tool vulnerabilities [26], unscrutinised data and data corruption.
  Automatic: Meta-data derived from external sources may contain malicious code, unscrutinised data, data corruption, poisoning attacks.
  Both: Attacks targeting the interface between annotation tool and ANN or targeting the functions involved in the import of the labelled data.
- Pre-processing unit – Implementation dependent, entry/exit point
  Operation on unscrutinised data, library vulnerabilities.
- Feature extraction – Implementation dependent, entry/exit point
  Operation on unscrutinised data, library vulnerabilities, database and import function vulnerabilities.
- Classifier – Exit point
  May impose threats to the overall application if data authenticity and access authorisation are not secured.
- Weights – Exit point (training); Entry point (shared weights → transfer learning)
  Authenticity of weights must be guaranteed.
  Access should be restricted to prevent theft or leakage.
  Database: database and import vulnerabilities apply.
  Volatile memory only: attack patterns against volatile memory apply.
  Shared weights: Transfer learning associated attack patterns such as NN trojans, unscrutinised data, poisoning attacks.

ANNs work with sensitive data assets. These must be protected to ensure the correctness and authenticity of the AI's output, as well as due to privacy considerations. The data assets found in AI systems utilising ANNs are:

- The data gathered itself;
- Labelled data [27] (backdoor triggers/poised data);
- Extracted features;
- Weights - Reports on volatile memory attacks exist [28], external weights obtained through model sharing may lead to trojan injections in NNs [29];
- Classification output.

Due to a lack of sufficient metrics for AI attack surfaces, it is difficult to derive a quantified and comparable assessment of the attack surface. However, it is observable that ANNs have a comparably large attack surface. The possibility of incorporating applications for the data gathering and annotation expand this attack surface even further. Overall, ANNs appear highly susceptible to a variety of cybersecurity attacks due to their complex nature of internal processing units and their frequent import/export of data requiring long-term storage.

When considering the security of the data assets, one must recall that the implementation is likely to allow concurrent processing. This implies the necessity for data management, which may either be solved through shared databases or complex merging strategies for distributed data. Both solutions imply specific attack vectors being utilisable – see section 3. A.

The application of transfer learning expands the attack surface even further, as another entry point within the ANN is established.

The above considerations provide insights into the efforts needed to secure applications utilising ANNs. The overall impression is that – without sufficient precautions being made – the attack surface of systems utilising ANNs is vast. Given the numerous reports of attack patterns directed against ANNs, this assessment appears reasonable.

B. CNNs

FIGURE 6. EXAMPLE APPLICATION UTILISING A CNN. CNNs may depict larger and more complex models as they do not have the common parameter space increase witnessed in ANNs. Pre-processing and feature extraction are performed by the CNN internally.

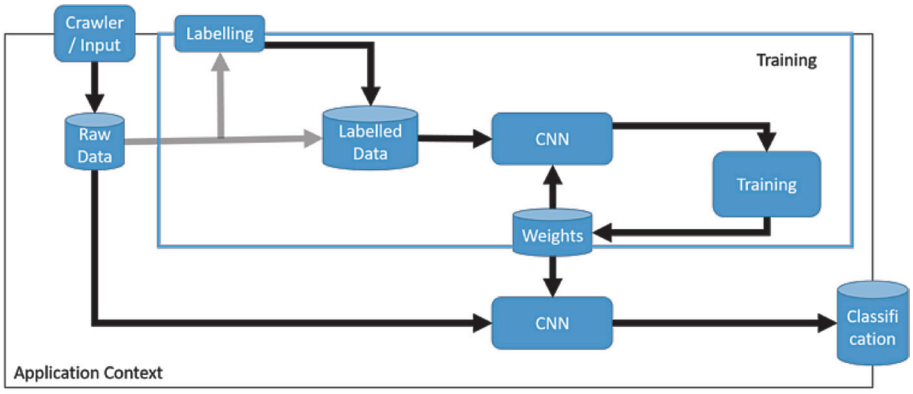


Figure 6 depicts an overview of a CNN in an abstract application context. The following attack surface points are identifiable:

- Crawler/Input – Entry point
  Susceptible to unscrutinised data, data corruption and poisoning attacks.
  Possibly susceptible to common (web) application vulnerabilities.
- Labelling – Entry point
  Manual annotation: Annotation tool vulnerabilities, unscrutinised data and data corruption.
  Automatic: Malicious meta-data, unscrutinised data, data corruption and poisoning attacks. Attacks directed against the interface between the annotation tool and the CNN (manual annotation) or against the data import of the labelled data from memory to CNN.
- Weights – Exit point (training); Entry point (shared weights, transfer learning)
  Authenticity of weights must be guaranteed.
  Access should be restricted to prevent theft or leakage.
  Database: database and import vulnerabilities apply.
  Volatile memory only: attack patterns against volatile memory apply.
  Shared weights: Due to the common utilisation of transfer learning, CNNs are particularly vulnerable to attack vectors utilising this method: Usage of externally trained weights for the CNN network may introduce logic bombs into the network [30]. This threat is hard to mitigate as it is difficult to anticipate the behaviour of CNNs based on the weights alone. The only option is to rigorously test the network with labelled data. Furthermore, NN trojans, unscrutinised data and poisoning attacks are plausible attack vectors.

CNNs work with sensitive data assets, these are:

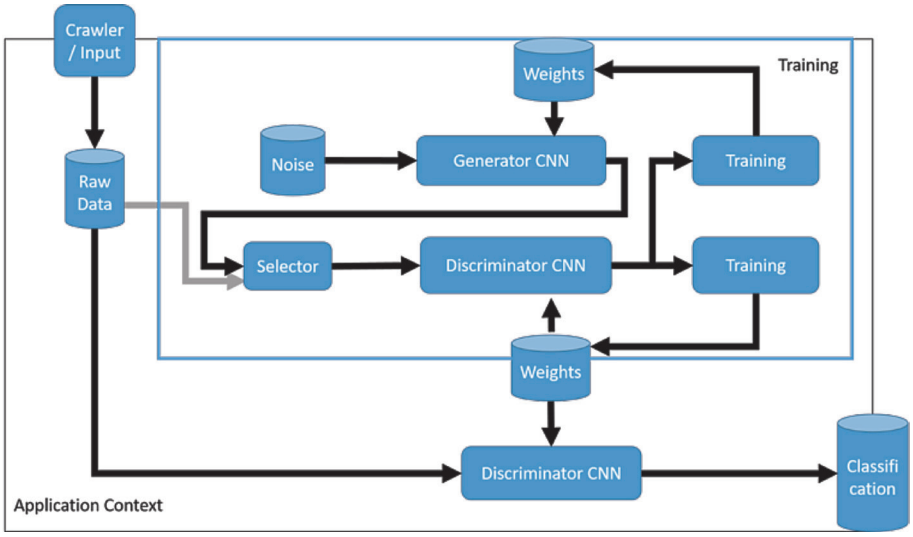
- The data gathered itself;
- Labelled data;
- Weights derived from training or through transfer learning;
- Classification results.

Within CNNs, the pre-processing and feature extraction are part of the network and not performed by separate application entities. Therefore, the data quality for CNN applications is of higher importance than for systems utilising ANNs.

In addition to the above, further attack vectors on CNNs have been reported, amongst others utilising evolutionary computing methods, evasion attacks and side-channel attacks on CNN FPGA accelerators [31].

C. GANs

FIGURE 7. AN EXEMPLARY GAN APPLICATION SYSTEM. The GAN is used to enhance the training of an already existing CNN (Discriminator CNN) for classification purposes. The Generator CNN creates additional training samples which are aimed to throw off the classification. The resulting Discriminator CNN after training is in general more robust against adversarial samples then the original one.



The attack surface is given by the systems entry/exit points, which are:

- Crawler/Input – Entry point
  See considerations in sections 4. A and B. However, for unconditional GANs such as the one shown in Figure 7, data integrity and authenticity is even more important, as no additional labels are used for the generative network. Therefore, all data points are equally important. Modification of the stochastic distribution of data may modify the behaviour of the whole GAN. The result is highly dependent on the used input data and appropriate training parameters [32].
- Weights – Exit point (training), entry point (training, shared weights, transfer learning)
  Within GANs, the weights may serve as exit and entry points.

Import/Export may be vulnerable to attacks on the interface or database used. Transfer learning is commonly used in GANs – implying GAN-based systems to be vulnerable to transfer learning attacks.

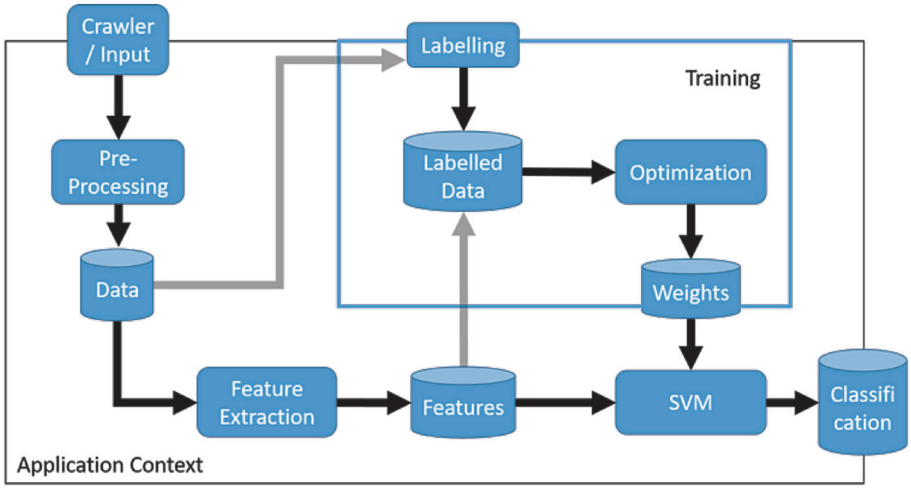
- Noise generator – (Hidden) Entry point
  The stochastic distribution of the random input used by the generator is crucial for the correct behaviour of the GAN. If the distribution is biased towards certain values, this will affect the training of both networks and may create blind spots as some data values are never generated and, therefore, the discriminator is not trained on them.
- Selector – (Hidden) Entry/exit point
  Attacks against the selector may yield a modification of the data passed. Furthermore, the selection process may be biased, yielding negative training outcomes due to an overrepresentation of real data (disabling the generator training) or an overrepresentation of artificial data (overfitting of the discriminator).
- Labelling – Entry point
  The shown unconditional GAN does not need labelled data, therefore this entry point is only present in conditional GANs. Similar to CNNs, modifying labels may negatively impact training of the discriminator. By changing labels of a specific class only, this class can be removed from the GAN altogether, preventing the generator from producing appropriate data points and also preventing the discriminator from classifying them.

Due to their composition of two CNNs operating in parallel, GANs have the same type of sensitive data assets as CNNs. Depending on the type of GAN (conditional or unconditional) labels may be present in the data (or not) and must be considered accordingly when defining the attack surface.

Most reported attacks on GANs try to reconstruct the used training data from the final model, which is called member inference attack [33]. These models can be used to generate adversarial attacks on other ML methods and also to protect them from such attacks [34].

D. SVMs

FIGURE 8. SVM APPLICATION SEPARATING LABELLED DATA INTO TWO CLASSES. Similarly to the ANNs discussed previously, pre-processing and feature extraction are performed separately from the training. The training is performed through mathematical optimisation. The SVM is executed after the training.



The following attack surface points are derivable:

- Crawler / Input – Entry point
  Unscrutinised data, data corruption and poisoning attacks.
  However, in contrast to ANNs, only a small fraction of the data defines the output. These are the support vectors identified during the training. Therefore, adversarial support vectors may heavily influence the resulting classification [35]. This type of poisoning attack is even possible in online learning environments where the SVM is continuously updated with new data [36]. Another approach uses poisoned data to prevent the training from converging through the introduction of artificially large training errors [37]. This can be used in online learning to prevent the system from updating the SVM.
- Weights – Entry point
  SVMs store a single weight per data point trained. Any data point that is not a support vector has a weight of zero. Weight modifications may therefore drastically change the output classification as it may alter the support vector identification. This allows for arbitrary output classification.

- Feature Storage – Entry/Exit point
  As the SVM is executed post-training and after the processing of the data input, it is dependent on accessing the data derived during these steps. Therefore, the feature storage is of particular importance to SVMs. An attack vector utilising this vulnerability is called the "label flip"-attack. It allows an attacker to change the label assigned to a support vector in order to change the final classification [38].
- Pre-processing and Feature Extraction – Entry points
  Data corruption and injection of malicious code in meta-data may enable an attacker to gain access to the system.

SVMs work on the following sensitive data assets:

- Raw data gathered;
- Pre-processed data;
- Features extracted;
- Labelled data;
- Weights derived – considered as the most important data points and features [39];
- Classification.

5. CONCLUSION AND OUTLOOK

Summarising the above findings and discussions, the combination of being inherently covert, their devastating impact on society and the wide unawareness of AI and ML vulnerabilities make attack vectors against these systems highly favourable for malicious cyber operators. Such attacks have already been witnessed and are being discussed in technical and academic communities but have not yet reached the public sphere, nor are application developers aware of the risk imposed by the utilisation of AI.

Despite the analyses presented in section 4, it remains difficult to provide a vulnerability hierarchy of the methods investigated regarding their susceptibility to cyber attacks. While some entry/exit points are easier to attack, others are only accessible with insider knowledge. The impact of the attack varies greatly with the data assets targeted and the specific method used. Using a preliminary approach to derive a quantifiable hierarchy based on the number of possible entry/exit points, one may observe that the number of entry/exit points is lowest in CNNs, followed by GANs and ANNs. SVMs have the same amount of identified entry/exit points as GANs. However, for AI systems, the mere number of entry/exit points is not a good

measure of the susceptibility of the technology investigated. It appears that each of the AI/ML methods investigated have specific high-value data assets, which make the system vulnerable through a combination of the data asset and a specific trait or process utilised. As an example, SVMs are highly sensitive to support vector manipulations, while GANs are exceptionally vulnerable to transfer learning attacks. The likelihood of successfully manipulating, destroying or obtaining these specific assets, traits or processes appears to give a more reliable assessment of the susceptibility than merely counting the overall number of access points. This is due to the fact that not all assets are equally important for the system to uphold its function, nor do all assets allow manipulation by an attacker or interact with the system.

In conclusion, it must be noted that AI systems are indeed susceptible to cyber attacks and that the utilisation of AI or ML methods increases any applications' vulnerability. This necessitates more sensitive use of AI and ML methods in security- or safety-sensitive applications.

Defining the attack surface of AI systems has provided information that requires further interpretation to derive the application specific risk of utilising AI/ML in the application context. Currently, only a few reports exist on attack surface metrics [40], and these are not specific to AI systems. We have seen that these systems cannot be analysed by solely investigating attack surfaces, but that the internal processing discloses particular weaknesses that are a result of the data assets used and the characteristics and processes of the methods used. Recent attacks against AI systems have shown that vulnerabilities are a result of the combination of particular AI architectures, the methods used, implementation decisions (data sharing, framework and library choices) as well as the data processing, storage and handling itself.

In order to enhance the security of AI systems, a common language to discuss the vulnerability of such systems must be installed. Furthermore, methods to reliably quantify systems' susceptibility to cyber attacks must be developed.

Policy considerations being driven by the AI community show that the need to harden AI systems against manipulations and attacks has been acknowledged within academic communities. Preliminary results from within the EU have been achieved by the Fraunhofer IAIS and the University of Bonn, who cooperated with the German Federal Office for Information Security to define a certification standard for AI, including security considerations. These results follow the EU AI HLEG and the EU AI Alliance working on the European Strategy on Artificial Intelligence.

Given the anticipated ubiquitous utilisation of AI and ML in applications over the next decade, the already existing diversity of attack vectors and the current inferiority of

countermeasures is alarming. The defence of AI systems is yet at its beginning and requires further investigation into the specific vulnerabilities of these systems [41]. Furthermore, knowledge of AI systems' vulnerabilities may become crucial to defend against cyber operations which are being carried out with the aid of AI. Such operations are currently described in modern disinformation campaigns, as well as in information and hybrid warfare with only limited countermeasures currently available. In the context of political challenges and the ongoing AI arms race, a profound knowledge of AI systems' vulnerabilities must be established to uphold cyber sovereignty.

REFERENCES

[1] Tom Simonite, "For Superpowers, Artificial Intelligence Fuels New Global Arms Race", 9 August 2017, https://www.wired.com/story/for-superpowers-artificial-intelligence-fuels-new-global-arms-race/; Catherine Clifford, "In the same way there was a nuclear arms race, there will be a race to build A.I., says tech exec", Interview with Hootsuite CEO Ryan Holmes on AI arms race, 29 September 2017, https://www.cnbc.com/2017/09/28/hootsuite-ceo-next-version-of-arms-race-will-be-a-race-to-build-ai.html.

[2] Steven Feldstein, "The Global Expansion of AI Surveillance", Carnegie Endowment for International Peace - Paper, 17 September 2019, https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847; Bruce Schneier, "AI Has Made Video Surveillance Automated and Terrifying", Motherboard - Tech by Vice, 13 June 2019, https://www.vice.com/en_us/article/bj93z5/ai-has-made-video-surveillance-automated-and-terrifying.

[3] Jun Wu, "Artificial Intelligence and The Trader", towardsdatascience.com, 28 May 2019, https://towardsdatascience.com/artificial-intelligence-and-the-trader-500745011f53; Mike Thomas, "How AI Trading Technology is Making Stock Market Investors Smarter — and Richer - AI Trading: 17 Companies Changing The Stock Market", builtin.com, 16 March 2019, https://builtin.com/artificial-intelligence/ai-trading-stock-market-tech.

[4] Sam Daley, "Surgical robots, new medicines and better care: 32 examples of AI in healthcare", builtin.com, 23 September 2019, https://builtin.com/artificial-intelligence/artificial-intelligence-healthcare.

[5] Sandra Ponce de Leon, Cognitive World, "The Role Of Smart Grids And AI In The Race To Zero Emissions", Forbes, 20 March 2019, https://www.forbes.com/sites/cognitiveworld/2019/03/20/the-role-of-smart-grids-and-ai-in-the-race-to-zero-emissions/#b5a97221c8e3.

[6] "Computer Vision (CV) dazzle" has been inspired from dazzle camouflage used by warships in World War I and involves make-up, haircut or infrared lights to distract automated facial recognition. Further reading: Elise Thomas, "How to hack your face to dodge the rise of facial recognition tech", Wired Magazine, 1 February 2019, https://www.wired.co.uk/article/avoid-facial-recognition-software; Samantha Cole, "This Trippy T-Shirt Makes You Invisible to AI", Vice Tech, 5 November 2019, https://www.vice.com/en_us/article/evj9bm/adversarial-design-shirt-makes-you-invisible-to-ai; Jonathan Vanian, "Why Google's Artificial Intelligence Confused a Turtle for a Rifle", fortune.com, 8 November 2017, https://fortune.com/2017/11/08/google-artificial-intelligence-turtle-rifle/.

[7] Assim Rais Siddiqui, "5 Security Measures for Verified Artificial Intelligence - Find out how to ensure a secure and trusted AI system for your business", business.com, 26 August 2019, https://www.business.com/articles/security-measures-verified-artificial-intelligence/.

[8] Valecia Maclin, "Solving the challenge of securing AI and machine learning systems", Microsoft Blog, 6 December 2019, https://blogs.microsoft.com/on-the-issues/2019/12/06/ai-machine-learning-security/.

[9] Keir Giles, Kim Hartmann, Munira Mustaffa, "The Role of Deepfakes in Malign Influence Campaigns", NATO StratCom COE, ISBN 978-9934-564-50-5, September 2019, https://www.stratcomcoe.org/role-deepfakes-malign-influence-campaigns.

[10] Stephanie Kampf, Mark Kelley, "A new 'arms race': How the U.S. military is spending millions to fight fake images", CBC.ca, 18 November 2018, https://www.cbc.ca/news/technology/fighting-fake-images-military-1.4905775.

[11] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, "You only look once: Unified, real-time object detection", in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779-788, https://doi.org/10.1109/CVPR.2016.91.

[12] Elie Bursztein, Security and Anti-Abuse Research Lead at Google, “Attacks against machine learning — an overview” Personal Site and Blog featuresing blog posts, publications and talks, May 2018, https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/.
[13] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, Seraphin B. Calo, “Analyzing Federated Learning through an Adversarial Lens”, *Proceedings of the 36th International Conference on Machine Learning*, PMLR 97:634-643, 2019, http://proceedings.mlr.press/v97/bhagoji19a.html.
[14] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, Thomas Ristenpart, “Stealing Machine Learning Models via Prediction APIs”, *Proceedings of the 25th USENIX Security Symposium*, August 2016, ISBN: 978-1-931971-32-4, https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_tramer.pdf; Itay Mosafi; Eli Omid David; Nathan S. Netanyahu, “Stealing Knowledge from Protected Deep Neural Networks Using Composite Unlabeled Data”, *Proceedings of 2019 International Joint Conference on Neural Networks (IJCNN)*, IEEE, July 2019, https://ieeexplore.ieee.org/abstract/document/8851798.
[15] Taylor Larsen, “Data leakage in healthcare machine learning”, healthcare.ai, obtained 7 January 2020, https://healthcare.ai/data-leakage-in-healthcare-machine-learning/; Jason Brownlee,“Data Leakage in Machine Learning”, machinelearningmastery.com, 2 August 2016, https://machinelearningmastery.com/data-leakage-machine-learning/.
[16] Yu Ji, Zixin Liu, Xing Hu, Peiqi Wang, Youhui Zhang, “Programmable Neural Network Trojan for Pre-Trained Feature Extractor”, arXiv.com, 23 January 2019, https://arxiv.org/abs/1901.07766v1; Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang, “Trojaning Attack on Neural Networks”, Purdue University - Department of Computer Science Technical Reports, Paper 1781, 2017, https://docs.lib.purdue.edu/cstech/1781.
[17] OWASP Foundation, “Attack Surface Analysis”, OWASP Cheatsheet Series, obtained 7 January 2020, https://cheatsheetseries.owasp.org/cheatsheets/Attack_Surface_Analysis_Cheat_Sheet.html.
[18] Lily Hay Newman, “Hacker Lexicon: What Is an Attack Surface?”, wired.com, 3 December 2017, https://www.wired.com/2017/03/hacker-lexicon-attack-surface/.
[19] Sven Herping, “Securing Artificial Intelligence – Part I”, October 2019, https://www.stiftung-nv.de/sites/default/files/securing_artificial_intelligence.pdf .
[20] Dana Neustadter, “Why AI Needs Security”, Synopsys Technical Bulletin, obtained 7 January 2020, https://www.synopsys.com/designware-ip/technical-bulletin/why-ai-needs-security-dwtb-q318.html; Alexander Polyakov, “AI Security and Adversarial Machine Learning 101”, towardsdatascience.com, 23 July 2019, https://towardsdatascience.com/ai-and-ml-security-101-6af8026675ff.
[21] Jeffrey Ding, “ChinAI #47: The Sensenet Data Leak - What Actually Happened”, 25 March 2019, https://chinai.substack.com/p/chinai-47-the-sensenet-data-leak.
[22] BBC Technology, “AI image recognition fooled by single pixel change”, 3 November 2017, https://www.bbc.com/news/technology-41845878.
[23] Eykholt, Kevin, et al. “Robust physical-world attacks on deep learning visual classification, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 1625-1634, https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00175.
[24] Savia Lobo, “VLC media player affected by a major vulnerability in a 3rd library, libebml; updating to the latest version may help”, hub.packtpub.com,25 July 2019, https://hub.packtpub.com/vlc-media-player-affected-by-a-major-vulnerability-in-a-3rd-library-libebml-updating-to-the-latest-version-may-help/: CVE-2019-13615 Details, NIST National Vulnerabilities Database, 16 July 2019, https://nvd.nist.gov/vuln/detail/CVE-2019-13615.
[25] OWASP Foundation, “Web Application Security Guidance”, obtained 8 January 2020, https://www.owasp.org/index.php/Web_Application_Security_Guidance; OWASP, “OWASP Top 10 Most Critical Web Application Security Risks”, OWASP Top Ten Project, obtained 8 January 2020, https://www.owasp.org/index.php/Category:OWASP_Top_Ten_Project.
[26] OWASP Foundation, “Application Security Verification Standard 4.0”, March 2019, https://www.owasp.org/images/d/d4/OWASP_Application_Security_Verification_Standard_4.0-en.pdf.
[27] Duke University Press Release, “Detecting backdoor attacks on artificial neural networks”, 23 December 2019, https://ece.duke.edu/about/news/detecting-backdoor-attacks-artificial-neural-networks.
[28] Adnan Siraj Rakin, Zhezhi He, Deliang Fan, “Bit-Flip Attack: Crushing Neural Network with Progressive Bit Search”, arXiv.com, 7 April 2019, https://arxiv.org/abs/1903.12269.
[29] Zhaoyuan Yang, Naresh Iyer, Johan Reimann, Nurali Virani, “Design of intentional backdoors in sequential models”, arXiv.com, 26 February 2019, https://arxiv.org/abs/1902.09972.
[30] Gu, Tianyu, Brendan Dolan-Gavitt, and Siddharth Garg, “Badnets: Identifying vulnerabilities in the machine learning model supply chain”, *arXiv preprint arXiv:1708.06733* (2017).

[31] Jiawei Su, Danilo Vasconcellos Vargas, Kouichi Sakurai, “Attacking convolutional neural network using differential evolution”, IPSJ Transactions on Computer Vision and Applications issue 11, 22 February 2019, https://link.springer.com/article/10.1186/s41074-019-0053-3; Ya-guan Qian, Dan-feng Ma, Bin Wang, Jun Pan, Jia-min Wang, Jian-hai Chen, Wu-jie Zhou, Jing-sheng Lei, “Spot Evasion Attacks: Adversarial Examples for License Plate Recognition Systems with Convolutional Neural Networks”, arXiv.com, 28 November 2019, https://arxiv.org/abs/1911.00927; Joao Gomes, “Adversarial Attacks and Defences for Convolutional Neural Networks”, medium.com, 16 January 2018, https://medium.com/onfido-tech/adversarial-attacks-and-defences-for-convolutional-neural-networks-66915ece52e7; Lingxiao Wei, Bo Luo, Yu Li, Yannan Liu, Qiang Xu, “I Know What You See: Power Side-Channel Attack on Convolutional Neural Network Accelerators”, in *ACSAC ‘18: Proceedings of the 34th Annual Computer Security Applications Conference*, 393–406, December 2018, https://dl.acm.org/doi/10.1145/3274694.3274696.

[32] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta and A. A. Bharath, “Generative Adversarial Networks: An Overview,” in IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 53-65, Jan. 2018.

[33] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro, “LOGAN: Membership inference attacks against generative models”, in *Proceedings on Privacy Enhancing Technologies 2019*, 1, pp. 133-152; Dingfan Chen, Ning Yu,Yang Zhang, Mario Fritz, “Gan-leaks: A taxonomy of membership inference attacks against gans”, arXiv preprint arXiv:1909.03935, 2019.

[34] Samangouei, Pouya, Maya Kabkab, and Rama Chellappa. “Defense-gan: Protecting classifiers against adversarial attacks using generative models”, arXiv preprint arXiv:1805.06605, 2018.

[35] Wang, Baoyao, Peidong Zhu, Yingwen Chen, Peng Xun, and Zhenyu Zhang, “False Data Injection Attack Based on Hyperplane Migration of Support Vector Machine in Transmission Network of the Smart Grid”, Symmetry 2018, 10(5), 165, https://doi.org/10.3390/sym10050165.

[36] Xiaojun Lin and Patrick P. K. Chan, “Causative attack to Incremental Support Vector Machine”, *Proceedings of 2014 International Conference on Machine Learning and Cybernetics*, IEEE, July 2014, https://doi.org/10.1109/ICMLC.2014.7009106.

[37] Battista Biggio, Blaine Nelson, and Pavel Laskov, “Poisoning Attacks against Support Vector Machines”, in *Proceedings of the 29th International Conference on Machine Learning*, 25 March 2013, https://arxiv.org/pdf/1206.6389.pdf.

[38] Han Xiao, Huang Xiao, and Claudia Eckert, “Adversarial Label Flips Attack on Support Vector Machines”, in *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI)*, August 2018; Ambra Demontis, Battista Biggio, Giorgio Fumera, Giorgio Giacinto, Fabio Roli, “Infinity-norm Support Vector Machines against Adversarial Label Contamination”, in *Proceedings of the 1st Italian Conference on Cybersecurity (ITASEC17)*, 2017, http://ceur-ws.org/Vol-1816/paper-11.pdf.

[39] Battista Biggio, Igino Corona, Blaine Nelson, Benjamin IP Rubinstein, Davide Maiorca, Giorgio Fumera, Giorgio Giacinto, Fabio Roli, “Security evaluation of support vector machines in adversarial environments”, Support Vector Machines Applications, pp. 105-153. Springer, Cham, 2014, https://arxiv.org/pdf/1401.7727.pdf.

[40] Pratyusa K. Manadhata, Jeannette M. Wing, “An Attack Surface Metric”, IEEE Transactions on Software Engineering (Volume: 37, Issue: 3), May-June 2011, https://doi.org/10.1109/TSE.2010.60.

[41] Adam Hadhazy, “Protecting smart machines from smart attacks”, Princeton Office of Engineering Communications, 14 October 2019, https://www.princeton.edu/news/2019/10/14/adversarial-machine-learning-artificial-intelligence-comes-new-types-attacks, quote from the text: “If machine learning is the software of the future, we’re at a very basic starting point for securing it” – Prateek Mittal, lead researcher and an associate professor in the Department of Electrical Engineering at Princeton.

2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Recent Developments in Cryptography

Lubjana Beshaj
Assistant Professor
Army Cyber Institute
United States Military Academy
West Point, New York,
United States of America
Lubjana.Beshaj@westpoint.edu

Andrew O. Hall
Associate Professor and Director
Army Cyber Institute
United States Military Academy
West Point, New York,
United States of America
Andrew.Hall@westpoint.edu

Abstract: In this short note, we briefly describe cryptosystems that are believed to be quantum-resistant and focus on isogeny-based cryptosystems. Recent SIDH (Supersingular Isogeny Diffie-Hellman) developments have focused on (2,2)-reducible Jacobians, where addition is executed via the Kummer surface. While elliptic curve isogenies are easy, explicit, and fast to compute thanks to Velús formulas, this is not the case for higher genus curves. The case of (2,2)-isogenies in genus 2 curves are an exception thanks to the work of Richelot. In addition, some explicit work has been completed in the case of (3,3) and (5,5)-isogenies, which are much more complicated than the case of Richelot isogenies. In this paper, we further investigate the case of (4,4)-reducible Jacobians and explicitly compute the locus $\mathcal{L}_4$.

Keywords: *quantum computing, post quantum cryptography, supersingular elliptic curves, Jacobian surfaces, isogenies, split Jacobians*

1. INTRODUCTION

Quantum computers are powerful machines that take a new approach to processing information and may lead to revolutionary breakthroughs in a variety of areas to include artificial intelligence, drug discovery, materials science, and optimization of complex man-made systems. While increased computational power, such as that offered by quantum computers, can be used for good, these advances do present a threat to public key cryptography. Public key cryptography, and cryptography in general, rely on computational hard or expensive problems. Problems that were extremely hard when only equipped with a pencil and paper are now easily solved with a classical computer. While hard problems for classical computing, like the discrete log problem, ensure the strength of today's current public key cryptography, new quantum algorithms can address these hard problems in polynomial time. Peter Shor, in his paper [37], provided an algorithm to solve the discrete log problem, demonstrating how to use a quantum computer to factor a positive odd integer. With the advent of these quantum algorithms, an adversary could efficiently break the universally adopted public-key cryptographic schemes (e.g. RSA, DSA and elliptic-curve cryptography).

In order to mitigate against this imminent threat, cryptographic schemes that are resistant to increased computing power offered by quantum computers have drawn great attention from both academia and industry. These schemes are collectively referred to as post-quantum cryptography (PQC). Whereas some cryptographic schemes will be rendered obsolete, several existing protocols, (e.g. current symmetric cryptography) do not need to be changed significantly to be considered quantum-resistant (i.e. post-quantum symmetric cryptography).

In April 2016, the National Institute of Science and Technology (NIST) initiated a process to solicit, evaluate, and standardize one or more quantum-resistant public-key cryptographic algorithms. They announced the release of NIST Interagency Report (NISTIR) 8105, a report on Post Quantum Cryptography (see [5] for more details). In this report, they explain the status of quantum computing and post-quantum cryptography, and outline a research plan for future work in these areas. In December 2016, NIST announced a formal call for proposals.

In the first round, 69 algorithms were submitted in response to the call for proposals and competition. Detailed information concerning these algorithms and the comments provided by the world-wide cryptography community are available on the NIST webpage (https://csrc.nist.gov/Projects/Post-Quantum-Cryptography).

As the latest step in the program to develop effective defenses and new standards, NIST has selected 26 of the 69 submitted cryptographic algorithms. There are 17

second round candidates for public-key encryption and key-establishment algorithms and 9 second round candidates for digital signatures. This second round will focus on evaluating submissions performance across a wide variety of systems and platforms as a variety of devices will require effective encryption.

After the completion of the second round of reviews, there still exists the possibility of an additional round of review before NIST announces the post-quantum algorithms that will supplement or replace the most vulnerable cryptosystems currently in use. The state of quantum computer development will determine the requirement for a third round of competition.

A tentative timeline made public by NIST will be given in the following table.

TABLE 1: NIST TIMELINE

Feb 24-26, 2016	NIST Presentation at PQCrypto 2016: Announcement and outline of NIST's Call for Submissions
April 28, 2016	NIST releases NISTIR 8105, Report on Post-Quantum Cryptography
Dec 20, 2016	Formal Call for Proposal
Nov 30, 2017	Deadline for submissions
Dec 4, 2017	NIST Presentation at AsiaCrypt 2017: The Ship Has Sailed: The NIST Post-Quantum Crypto "Competition"
Dec 21, 2017	Round 1 algorithms announced (69 submissions accepted as "complete and proper")
Apr 11, 2018	NIST Presentation at PQCrypto 2018: Let's Get Ready to Rumble – The NIST PQC "Competition"
April 11-13, 2018	First PQC Standardization Conference Submitter's Presentations
2018/2019	Round 2 begins
August 2019	Second PQC Standardization Conference
2020/2021	Round 3 begins or select algorithms
2022/2024	Draft Standards Available

The new algorithms rely on several cryptographic schemes that are believed to be post-quantum-resistant and include the following:

1. Code-based cryptography;
2. Multivariate Cryptography;
3. Lattice-based Cryptography;
4. Hash-based Cryptography;
5. Isogeny-based Cryptography.

Each of these cryptographic schemes has advantages and disadvantages, and the algorithms vary in both their performance measures and maturity. In this paper, we will focus on isogeny-based cryptography.

Supersingular isogeny-based cryptography is one of the more recent advances based on the arithmetic of elliptic curves. In 2011, Jao and De Feo proposed Supersingular Isogeny Diffie-Hellman (SIDH) as a key exchange protocol that would offer post-quantum security. Isogeny-based algorithms rely on the structure of large isogeny graphs, and the cryptographically interesting properties of these graphs are tied to their expansion properties.

In recent developments in supersingular isogeny-based cryptography (SIDH), Costello [8] focuses on (2,2) reducible Jacobians, where addition is executed via Kummer surfaces. More importantly, it seems that the most interesting case is when E_1 is isogenous to E_2. In this case, as the decomposition of the Abelian varieties is determined up to isogeny, the 2-dimensional Jacobian is isogenous to E^2. There are several interesting questions that arise when we consider such Jacobians over the finite field $\mathbb{F}_p$.

The space of genus 2 curves with (n,n) reducible Jacobians, for which $n=2$ or where n is odd, is a 2-dimensional irreducible locus $\mathcal{L}_n$ in the moduli space of curves $\mathcal{M}_2$. For $n=2$, this is the well known locus of curves with extra involutions [23], [24], [35]. In the cases where n is odd, these spaces were computed for the first time in [32], [34], [22].

If E_1 and E_2 are N-isogenous then their j-invariants j_1 and j_2 satisfy the equation of the modular curve $X_0(N)$, say $\mathcal{S}_N := \phi_N(j_1,j_2)=0$. Such a curve can be embedded in $\mathcal{M}_2$. An interesting problem to consider is the study of the intersection between $\mathcal{L}_n$ and $\mathcal{S}_N$ for given n and N. More precisely, for any number field K determines the number of K-rational points of this intersection. For the case when $n=2,3$ this was done in [3]. The case when $n=4$ is more complicated since the locus $\mathcal{L}_4$ is not explicitly computed. The focus of this paper is to compute the locus $\mathcal{L}_4$ and then further inverstigate when the

two elliptic components of the (n,n) reducible 2-dimensional Jacobians are isogenous to each other when $n=4$ and $N=2,3,5,7\ldots..$

The remainder of this paper is organized as follows. First we provide an overivew of quantum computing and briefly explain Shor and Grover's algorithms. In Section 3, we describe each of the cryptosystems mentioned above. Also, we further explain the small changes that should be made to the AES algorithm to allow for its continued use and to ensure its ability to resist exploitation by quantum computers. We briefly explain supersingular isogeny Diffie-Hellman key exchange, and finally explore (n,n)-split Jacobians and compute the locus $\mathcal{L}_4$.

2. QUANTUM COMPUTING

A classical computer has registers that are made up of bits, whereas a quantum computer has a single quantum register that is made up of qubits. Given q classical bits, their state is a binary string in $\{0,1\}^q$, which is a q-dimensional space. Whereas, a q-qubit quantum register is a 2^q -dimensional space. Hence, the dimension of the state space of a quantum computer grows exponentially while that of a classical computer grows linearly. Furthermore, the amount of information stored in a q-qubit quantum register is enormous compared with a classical q-bit computer. However, accessing the information stored in a quantum computer is not as easy as in a classical computer. Information on the quantum state is only gathered through a measurement gate.

One of the main questions regarding quantum computers is the type of algorithms that can be implemented on a quantum computer once they are fielded. There are three known algorithms that can be implemented on a quantum computer: Shor's, Grover's and Simon's algorithms.

In 1994, Peter Shor came up with a quantum algorithm that calculates the prime factors of a large number vastly more efficiently than a classical computer. This poses a threat to all modern cryptographic schemes that rely on the difficulty of factoring prime numbers. More generally, this algorithm poses a threat to all crypto-systems that rely on the difficulty of the discrete logarithm problem.

However, Shor's algorithm's efficiency and power relies on a quantum computer with a large number of quantum bits. It should be noted that Shor's algorithm is only partially executed on a quantum computer. While many have attempted to implement Shor's algorithm on various quantum systems, none have been successful in doing so with more than a few quantum bits or in a scalable way.

Grover's algorithm performs a search over an unordered set of $N=2^n$ items to find the unique element that satisfies some condition. Grover's algorithm performs the search on a quantum computer which is a quadratic speedup ($O(\sqrt{N})$) compared to the best classical algorithm ($O(N)$), i.e. a speedup on the brute force attack. In order to achieve such a speedup, Grover relies on the quantum superposition of states.

It has been shown that applying Grover's algorithm to break a symmetric key algorithm by brute force requires a time roughly $2^{n/2}$, compared to 2^n in the classical case. Hence the symmetric key lengths are halved, i.e. AES 256 would provide the same security level against an attack using Grover's algorithm as AES 128 would provide against a classical attack. Hence, as long as the best-known attack on AES is the brute force attack, we can classify AES as quantum-resistant.

Post-quantum symmetric cryptography does not need to be changed significantly from current symmetric cryptography other than by increasing current security levels. The AES algorithm with appropriate key length will be able to resist attacks launched from quantum computers.

3. POST-QUANTUM CRYPTOGRAPHY

In this section, we describe shortly different cryptosystems that are believed to be quantum-resistant. For more details, see [5] and the NIST webpage on post-quantum cryptosystems.

A. Code-based Cryptography

Code-based cryptosystems are among the most promising candidates to replace quantum-vulnerable primitives such as the Diffie-Hellman key exchange, the Rivest-Shamir-Adleman (RSA), and ElGamal cryptosystems. One of the problems for which no known polynomial time algorithm on a quantum computer exists is the decoding of a general linear code. Conservative and well-understood choices for code-based cryptography are the McEliece cryptosystem [25] and its dual variant by Niederreiter [27] using binary Goppa codes.

B. Multivariate Cryptography

Another potential candidate for PQC is multivariate cryptography. Multivariate cryptography relies on the difficulty of solving a system of m polynomial equations in n variables over a finite field. The complexity of solving a multivariate polynomial system (MP problem) or a multivariate quadratic system (MQ problem) where coefficients of the monomials are independently and uniformly distributed (i.e. random) is well-known to be NP-hard.

An arbitrary MP system can be transformed into an equivalent MQ system by substituting monomials of degree larger than two with new variables and introducing extra equations to the system. Furthermore, a polynomial system over any extension field $\mathbb{F}^{2n}$ can be reduced into an equivalent system over $\mathbb{F}^{2}$ using a Weil descent.

While there have been some proposals for multivariate encryption schemes, multivariate cryptography has historically been more successfully employed as an approach to signatures.

C. Lattice-based Cryptography

A lattice is an infinite arrangement of regularly spaced points, and can be generated as the set of all linear combinations of m independent vectors in $\mathbb{R}^n$, called a basis. Cryptosystems based on lattice problems have received renewed interest. Lattice-based cryptography starts with the work of Ajtai [1] and uses hard problems on lattices as the foundation of secure cryptographic constructions. Exciting new applications (such as fully homomorphic encryption, code obfuscation, and attribute-based encryption) have been made possible using lattice-based cryptography.

Lattice-based cryptographic constructions are mainly based on two well-known problems: the Small Integer Solution problem (SIS) and its Inhomogeneous variant (ISIS) [1], and the Learning With Errors problem (LWE) introduced by Regev [29]. Structured variants of the LWE and SIS problems were proposed [39], called Ring-SIS and Ring-LWE. These problems are preferred in practice since they enjoy smaller storage and faster operations. These two problems can be used to construct many basic cryptographic primitives such as PKE (adapting the schemes from [29]) and signatures [10], [11], [21].

D. Hash-based Cryptography

Cryptographic hash functions are one of the central primitives in cryptography. They are used virtually everywhere: as cryptographically secure checksums to verify the integrity of software or data packages; as building block in security protocols, including TLS, SSH, IPSEC; as part of any efficient variable-input-length signature scheme; to build fully-fledged hash-based signature schemes; and in transformations for CCA-secure encryption.

While all widely deployed means of public-key cryptography may be threatened by the rise of quantum computers, hash functions are believed to be only mildly affected. The reason for this is two-fold. On the one hand, generic quantum attacks achieve at most a square-root speed up compared to their pre-quantum counterparts and can be proven asymptotically optimal [15], [41]. On the other hand, no dedicated quantum

attacks on any specific hash function perform better than generic quantum attacks (except, of course, for hash functions based on number theory, e.g., VSH [6]).

E. Isogeny-based Cryptography

Supersingular isogeny-based cryptography is one of the more recent families of post-quantum proposals. Ever since their introduction to public-key cryptography by Miller [26] and Koblitz [18], elliptic curves have been of interest to the cryptographic community. By using the group of points on an appropriately chosen elliptic curve where the discrete logarithm problem is assumed to be hard, many standard protocols can be instantiated. The efficiency of these curve-based algorithms is largely determined by the scalar multiplication routine, and as a result extensive research has gone into optimizing this operation.

In 2011, Jao and De Feo [17] proposed supersingular isogeny Diffie-Hellman as a key exchange protocol offering post-quantum security.

4. ISOGENY-BASED SUPERSINGULAR ELLIPTIC CURVE CRYPTOGRAPHY

In this section, we will give a brief overview on supersingular isogeny-based cryptography and explain the quantum-resistant supersingular Diffie-Hellman key exchange scheme. Most of the material presented in this section can be found in [2, 4, 7, 12].

A. Isogenies of Elliptic Curves

Let E and E' be elliptic curves defined over field K. An isogeny $\phi{:}E{\to}E'$ is an algebraic morphism satisfying $\phi(\infty){=}\infty$. The degree of the isogeny is its degree as an algebraic map. The endomorphism ring $End\ (E)$ is the set of isogenies from E to itself, together with the constant morphism. This set forms a ring under point-wise addition and composition.

When K is a finite field, the rank of $\mathbb{E}nd(E)$ as a $\mathbb{Z}$-module is either 2 or 4. We say E is *supersingular* if the rank is 4, and ordinary otherwise. A supersingular curve cannot be isogenous to an ordinary curve.

Supersingular curves are all defined over $\mathbb{F}_{p2}$, and for every prime $l\gamma p$ there exist $l{+}1$ isogenies (counting multiplicities) of degree l originating from any given such supersingular curve. Given an elliptic curve E and a finite group G of E, there is up to isomorphism a unique isogeny $E{\to}E'$ having kernel G, [38]. Hence we can identify an isogeny by specifying its kernel, and conversely given a kernel subgroup the

corresponding isogeny can be found using Vélu's formulas, see [40]. Two elliptic curves are called *isogenous* if there exists an isogeny between them.

B. Supersingular Isogeny Diffie-Hellman Key Exchange

In this section, we present briefly a key exchange protocol using supersingular elliptic curves; see [12] for a more complete description of this protocol as well as zero-knowledge proof of identity and a public-key encryption based on supersingular isogenies.

This protocol requires supersingular curves of smooth order. Fix $\mathbb{F}_q=\mathbb{F}_{p2}$, where $p = l_A^{e_A} l_B^{e_B} \cdot f \pm 1$ and l_A, l_B are small primes, and f is a cofactor such that p is prime. Construct a supersingular elliptic curve E defined over $\mathbb{F}_q$ of cardinality $(l_A^{e_A} l_B^{e_B} \cdot f)^2$. By construction, $E[l_A^{e_A}]$ is $\mathbb{F}_q$-rational and contains $l_A^{e_A-1}(l_A+1)$ cyclic subgroups of order $l_A^{e_A}$, each defining a different isogeny; the analogous statement holds for $E[l_B^{e_B}]$.

More precisely, the supersingular isogeny Diffie-Hellman key exchange follows this algorithm. Pick as the public parameters a supersingular elliptic curve E over $\mathbb{F}_{p2}$, and bases $\{P_A,Q_A\}$ and $\{P_B,Q_B\}$ which generate respectively $E[\ell_A^{e_A}] = \langle P_A, Q_A\rangle$, and $E[\ell_B^{e_B}] = \langle P_B, Q_B\rangle$. Then Alice chooses two random numbers $m_A,n_A\in\mathbb{Z}$ not both divisible by l_A, and computes an isogeny $\alpha:E\to E/\langle A\rangle$ with kernel $\langle A\rangle=\langle [m_A]P_A+[n_A]Q_A\rangle$. Alice computes also $\alpha(P_B)$ and $\alpha(Q_B)$ and then sends them to Bob together with E_A.

Bob on the other side chooses two random numbers $m_B,n_B\in\mathbb{Z}$ not both divisible by l_B, and computes an isogeny $\beta:E\to E/\langle B\rangle$ with kernel $\langle B\rangle=\langle [m_B]P_AB+[n_B]Q_B\rangle$ as well as $\beta(P_A)$ and $\beta(Q_A)$ and then sends them to Alice.

Upon receipt of the respective information, both parties can compute the secret shared key. Alice computes $E/\langle A,B\rangle=E_B/\langle\beta(A)\rangle$ and $\langle\beta(A)\rangle=\langle [m_A]\beta(P_A)+[n_A]\beta(Q_A)\rangle$ and Bob similarly computes $E/\langle A,B\rangle=E_A/\langle\alpha(B)\rangle$ where $\langle\alpha(A)\rangle=\langle [m_B]\alpha(P_B)+[n_B]\alpha(Q_B)\rangle$ so that they have the shared secret key $E/\langle A,B\rangle$. This is summarised in the following table 2.

Given two elliptic curves E, E' over a finite field, isogenous of known degree d, finding an isogeny $\phi:E\to E'$ of degree d is a notoriously difficult problem for which only algorithms exponential in log $\#E$ are known in general.

In [9] they give a precise formulation of the necessary computational assumptions (of supersingular isogeny Diffie-Hellman key exchange, zero-knowledge proof of identity, and a public-key encryption based on supersingular isogenies) along with a discussion of their validity, and prove the security of these protocols under those assumptions.

However, in recent developments in supersingular isogeny-based cryptography (SIDH), Costello [8] focuses on (2,2) reducible Jacobians. As pointed out by Costello in the last paragraph of [8]: *"One hope in this direction is the possibility of pushing odd degree l-isogeny maps from the elliptic curve setting to the Kummer setting. This was difficult in the case of 2-isogenies because the maps themselves are (2, 2)-isogenies, but in the case of odd degree isogenies there is nothing obvious preventing this approach."*

TABLE 2: SUPERSINGULAR ISOGENY DIFFIE-HELLMAN KEY EXCHANGE ALGORITHM

Alice	Bob
Pick $k_{PrA} = \langle A \rangle = \langle [m_A]P_A + [n_A]Q_A \rangle$	Pick $k_{PrB} = \langle B \rangle = \langle [m_B]P_B + [n_B]Q_B \rangle$
Comp. secret isogeny	Comp. secret isogeny
$\alpha : E \to E_A = E/\langle A \rangle$	$\beta : E \to E_B = E/\langle B \rangle$
Send $E_A, \alpha(P_B), \alpha(Q_B)$ ⟶	to Bob
to Alice ⟵	Send $E_B, \beta(P_A), \beta(Q_A)$
Secret shared key: Compute $E/\langle A, B \rangle = E_B/\langle \beta(A) \rangle$ $\langle \beta(A) \rangle = \langle [m_A]\beta(P_A) + [n_A]\beta(Q_A) \rangle$	Secret shared key: Compute $E/\langle A, B \rangle = E_A/\langle \alpha(B) \rangle$ $\langle \alpha(B) \rangle = \langle [m_B]\alpha(P_B) + [n_B]\alpha(Q_B) \rangle$

In the upcoming sections, we focus on n,n-reducible Jacobians, and more precisely when n=4.

5. ISOGENOUS COMPONENTS OF JACOBIAN SURFACES

An Abelian variety defined over k is an absolutely irreducible projective variety defined over k, which is a group scheme. We will denote an Abelian variety defined over a field k by $\mathbb{A}_k$ or simply $\mathbb{A}$. A morphism from the Abelian variety $\mathbb{A}_1$ to the Abelian variety $\mathbb{A}_2$ is a homomorphism if and only if it maps the identity element of $\mathbb{A}_1$ to the identity element of $\mathbb{A}_2$.

An Abelian variety over a field k is called simple if it has no proper non-zero Abelian subvariety over k. It is called *absolutely simple* (*or geometrically simple*) if it is simple over the algebraic closure of k. An Abelian variety of dimension 1 is called an *elliptic curve*.

A homomorphism f:$\mathbb{A}\to\mathcal{H}$ is called an isogeny if $Imgf=\mathcal{H}$ and $\ker f$ is a finite group scheme. If an isogeny $\mathbb{A}\to\mathcal{H}$ exists, we say that $\mathbb{A}$ and $\mathcal{H}$ are isogenous. This relation is symmetric. The degree of an isogeny f:$\mathbb{A}\to\mathcal{H}$ is the degree of the function field extension $\deg f:=[k(\mathbb{A}):f\star k(\mathcal{H})]$. It is equal to the order of the group scheme $\ker(f)$, which is, by definition, the scheme theoretical inverse image $f^{-1}(\{0_{\mathbb{A}}\})$.

The group of $\bar{k}$-rational points has order $\#(\ker f)(\bar{k})=[k(A):f^{\star}k(B)]^{sep}$, where $[k(A):f^{\star}k(B)]^{sep}$ is the degree of the maximally separable extension in $k(A)/f^{\star}k(\mathcal{H})$. We say that f is a *separable isogeny* if and only if $\#\ker f(\bar{k})=\deg f$.

For any Abelian variety $\mathbb{A}/k$ there is a one to one correspondence between the finite subgroup schemes $H\leq\mathbb{A}$ and isogenies f:$\mathbb{A}\to\mathcal{H}$, where $\mathcal{H}$ is determined up to isomorphism. Moreover, $H=\ker f$ and $\mathcal{H}=\mathbb{A}/H$. f is separable if and only if K is étale, and then $\deg f=\#H(\bar{k})$. The following is often called the fundamental theorem of Abelian varieties. Let $\mathbb{A}$ be an Abelian variety. Then $\mathbb{A}$ is isogenous to $\mathbb{A}_1^{n_1}\times\mathbb{A}_2^{n_2}\times\ldots\times\mathbb{A}_r^{n_r}$, where (up to permutation of the factors) $\mathbb{A}_i$, for $i=1,\ldots,r$ are simple, non-isogenous, Abelian varieties. Moreover, up to permutations, the factors $\mathbb{A}_i^{n_i}$ are uniquely determined up to isogenies.

When $k=\bar{k}$, then let f be a non-zero isogeny of $\mathbb{A}$. Its kernel $\ker f$ is a subgroup scheme of $\mathbb{A}$. It contains $0_{\mathbb{A}}$ and so its connected component, which is, by definition, an Abelian variety.

A. Jacobian Surfaces

Abelian varieties of dimension 2 are often called Abelian (algebraic) surfaces. We focus on Abelian surfaces which are Jacobian varieties. Let $\mathcal{X}$ be a genus 2 curve defined over a field k. Then its gonality is $\gamma_x=2$. Hence, genus 2 curves are hyperelliptic and we denote the hyperelliptic projection by π:$\mathcal{X}\to\mathbb{P}^1$. By the Hurwitz's formula, this covering has $r=6$ branch points which are images of the Weierstrass points of $\mathcal{X}$. The moduli space has dimension $r-3=3$.

The arithmetic of the moduli space of genus two curves was studied by Igusa in his seminal paper [16] expanding on the work of Clebsch, Bolza, and others. Arithmetic invariants by J_2,J_4,J_6,J_8,J_{10} determine uniquely the isomorphism class of a genus two curve. Two genus two curves $\mathcal{X}$ and $\mathcal{X}'$ are isomorphic over $\bar{k}$ if and only if there exists $\lambda\in\bar{k}^{\star}$ such that $J_{2i}(\mathcal{X})=\lambda^{2i}J_{2i}(\mathcal{X}')$, for $i=1,\ldots,5$. If chark $\neq2$ then the invariant J_8 is not needed.

From now on we assume char $k\neq2$. Then $\mathcal{X}$ has an affine Weierstrass equation

$$y^2=f(x)=a_6x^6+\cdots+a_1x+a_0, \tag{1}$$

over $\bar{k}$, with discriminant $\Delta_f = J_{10} \neq 0$. The moduli space $\mathcal{M}_2$ of genus 2 curves, via the Torelli morphism, can be identified with the moduli space of the principally polarized abelian surfaces $\mathbb{A}_2$ which are not products of elliptic curves. Its compactification $\mathbb{A}_2^\star$ is the weighted projective space $\mathbb{WP}^3_{(2,4,6,10)}(k)$ via the Igusa invariants J_2, J_4, J_6, J_{10}. Hence, $A_2 \cong \mathbb{WP}^3_{(2,4,6,10)}(k) \setminus \{J_{10}=0\}$. Given a moduli point $\mathfrak{p} \in \mathcal{M}_2$, we can recover the equation of the corresponding curve over a minimal field of definition following [23].

It is well known that a map of algebraic curves $f: X \to Y$ induces maps between their Jacobians $f^*: \mathbb{J}ac(Y) \to \mathbb{J}ac(X)$ and $f_*: \mathbb{J}ac(X) \to \mathbb{J}ac(Y)$. When f is maximal then f^* is injective and $\ker(f^*)$ is connected; see [31] for more details.

Let $\mathcal{X}$ be a genus 2 curve and $\psi_1: \mathcal{X} \longrightarrow E_1$ be a degree n maximal covering from $\mathcal{X}$ to an elliptic curve E_1. Then $\psi_1^*: E_1 \to \mathbb{J}ac(\mathcal{X})$ is injective and the kernel of $\psi_{1,*}: \mathbb{J}ac(\mathcal{X}) \to E_1$ is an elliptic curve, which we denote by E_2. For a fixed Weierstrass point $P \in \mathcal{X}$, we can embed $\mathcal{X}$ to its Jacobian via

$$\begin{aligned} &i_p: \mathcal{X} \longrightarrow \mathbb{J}ac(\mathcal{X}) \\ &x \to [(x) - (P)] \end{aligned} \tag{2}$$

Let $g: E_2 \to \mathbb{J}ac(\mathcal{X})$ be the natural embedding of E_2 in $\mathbb{J}ac(\mathcal{X})$, then there exists $g^*: \mathbb{J}ac(\mathcal{X}) \to E_2$. Define $\psi_2 = g^* \circ i_p: \mathcal{X} \to E_2$. So we have the following exact sequence

$$0 \to E_2 g \longrightarrow \mathbb{J}ac(\mathcal{X}) \psi_{1,*} \longrightarrow E_1 \to 0. \tag{3}$$

The dual sequence is also exact $0 \to E_1 \psi_1^* \longrightarrow \mathbb{J}ac(\mathcal{X}) g^* \longrightarrow E_2 \to 0$.
If $\deg(\psi_1) = 2$ or it is an odd number, then the maximal covering $\psi_2: \mathcal{X} \to E_2$ is unique (up to isomorphism of elliptic curves). The Hurwitz space $\mathcal{H}_\sigma$ of such covers is embedded as a subvariety of the moduli space of genus two curves $\mathcal{M}_2$; see [34] for details. It is a 2-dimensional subvariety of $\mathcal{M}_2$ which we denote using $\mathcal{L}_n$. An explicit equation for $\mathcal{L}_n$, in terms of the arithmetic invariants of genus 2 curves, can be found in [35] or [23] for $n=2$, in [34] for $n=3$, and in [22] for $n=5$. From now on, we will say that a genus 2 curve $\mathcal{X}$ has an (n,n)-decomposable Jacobian if $\mathcal{X}$ is as above and the elliptic curves E_i, $i=1,2$ are called the components of $\mathbb{J}ac(\mathcal{X})$.

For every $D := J_{10} > 0$ there is a Humbert hypersurface H_D in $\mathcal{M}_2$ which parametrizes curves $\mathcal{X}$ whose Jacobians admit an optimal action on $\mathcal{O}_D$; see [14]. Points on H_{n^2} parametrize curves whose Jacobian admits an (n,n)-isogeny to a product of two elliptic curves. Such curves are the main focus of our study. In [20, Prop. 2.14] the authors prove that $\mathbb{J}ac(\mathcal{X})$ is a geometrically simple Abelian variety if and only if it is not (n,n)-decomposable for some $n > 1$.

6. (N,N) REDUCIBLE JACOBIANS SURFACES

Genus 2 curves with (n,n)-decomposable Jacobians are the most studied type of genus 2 curves due to work of Jacobi, Hermite, et al. They provide examples of genus two curves with a large Mordell-Weil rank of the Jacobian, many rational points, nice examples of descent [33], etc. Such curves have received new attention lately due to interest in their use on cryptographic applications and their suggested use on post-quantum crypto-systems and the random self-reducibility of discrete logarithm problem; see [8]. A detailed account of applications of such curves in cryptography is provided in [13].

Let $\mathcal{X}$ be a genus 2 curve defined over an algebraically closed field k, $chark=0$, K the function field of $\mathcal{X}$, and $\psi_1:\mathcal{X}\longrightarrow E_1$ a degree n covering from $\mathcal{X}$ to an elliptic curve E; see [31] for the basic definitions. The covering $\psi_1:\mathcal{X}\longrightarrow E$ is called a *maximal covering* if it does not factor through a nontrivial isogeny. We call E a *degree n elliptic subcover* of $\mathcal{X}$. Degree n elliptic subcovers occur in pairs, say (E_1,E_2). It is well known that there is an isogeny of degree n^2 between the Jacobian $\mathbb{J}ac(\mathcal{X})$ and the product $E_1\times E_2$. Such curve $\mathcal{X}$ is said to have (n,n)-decomposable (or (n,n)-split) Jacobian. The focus of this paper is on isogenies among the elliptic curves E_1 and E_2.

The locus of genus 2 curves $\mathcal{X}$ with (n,n)-decomposable Jacobian it is denoted by $\mathcal{L}_n$. When $n=2$ or n an odd integer, $\mathcal{L}_n$ is a 2-dimensional algebraic subvariety of the moduli space $\mathcal{M}_2$ of genus two curves; see [31] for details. Hence, we can get an explicit equation of $\mathcal{L}_n$ in terms of the Igusa invariants J_2,J_4,J_6,J_{10}; see [35] for $\mathcal{L}_2$, [34] for $\mathcal{L}_3$, [36] for $\mathcal{L}_4$, and [22] for $\mathcal{L}_5$. There is a more recent paper on the subject [19] where results of [22, 34] are confirmed and equations for $n>5$ are studied.

A. Computing the Locus $\mathcal{L}_4$ in $\mathcal{M}_2$

When $\deg(\phi)=4$ to compute the locus $\mathcal{L}_4(\sigma)$ one has to consider two cases. There is one generic case and one degenerate case with possible ramification structures:

1. $(2,2,2,2^2,2)$ (generic)
2. $(2,2,2,4)$ (degenerate)

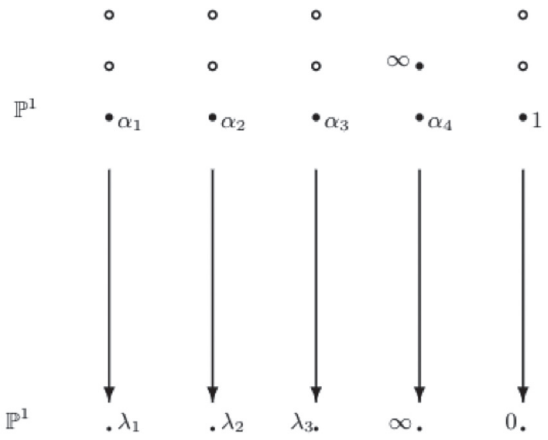
In this paper, we will focus only on the generic case. For a complete treatment of the degenerate case see [28, 36].

B. Non-degenerate Case

Let $\psi:C\longrightarrow E$ be a covering of degree 4, where C is a genus 2 curve and E is an elliptic curve. Let ϕ be the Frey-Kani covering with $\deg(\phi)=4$ such that $\phi(1)=0$, $\phi(\infty)=\infty$,

$\phi(p)=\infty$ and the roots of $f(x)=x^2+ax+b$ be in the fiber of 0. In the following figure, bullets (resp., circles) represent places of ramification index 2 (resp., 1).

FIGURE 2: DEGREE 4 COVERING FOR GENERIC CASE



Then the cover can be given by

$$\phi(x) = \frac{k(x-1)^2(x^2 + ax + b)}{(x-\alpha_4)^2}. \tag{4}$$

Let $\lambda_1,\lambda_2,\lambda_3$ and ∞ be the Weierstrass points of E. Then

$$\begin{cases} \phi(x) - \lambda_1 = k\dfrac{(x-\alpha_1)^2(x^2 - a_1x + b_1)}{(x-\alpha_4)^2} \\ \phi(x) - \lambda_2 = k\dfrac{(x-\alpha_2)^2(x^2 - a_2x + b_2)}{(x-\alpha_4)^2} \\ \phi(x) - \lambda_3 = k\dfrac{(x-\alpha_3)^2(x^2 - a_3x + b_3)}{(x-\alpha_4)^2} \end{cases}$$

Next, let $\lambda_1,\lambda_2,\lambda_3$ and 0 be the Weierstrass points of E. Then

$$\begin{cases} \phi(x) - \lambda_1 = k\,x\,(x-\alpha_1)^2(x^2 - a_1x + b_1) \\ \phi(x) - \lambda_2 = k\,x\,(x-\alpha_2)^2(x^2 - a_2x + b_2) \\ \phi(x) - \lambda_3 = k\,x\,(x-\alpha_3)^2(x^2 - a_3x + b_3) \end{cases}$$

By clearing the denomirators and equaling the coefficients of quartics to zero, we get a system of equations in terms of parameters a,b,a_1,b_1,a_2,b_2,a_3,b_3,$\alpha 1$,...$\alpha 4$, λ_1,λ_2,λ_3,k. We solve this equation to get

$$\begin{cases} \alpha_1 = -3\,a + 2 + A \\ \alpha_2 = -3\,a + 2 + A \\ \alpha_3 = -3\,a + 2 + A \\ \lambda_1 = aA^{3/2} - 27\,a^4 + 18\,a^2A - 72\,a^3 + 144\,a^2b + 8\,aA - 64\,bA - 56\,a^2 + 320\,ba \\ \qquad -128\,b^2 + 8\,A + 32\,a + 320\,b + 16 \\ \lambda_2 = aA^{3/2} - 27\,a^4 + 18\,a^2A - 72\,a^3 + 144\,a^2b + 8\,aA - 64\,bA - 56\,a^2 + 320\,ba \\ \qquad -128\,b^2 + 8\,A + 32\,a + 320\,b + 16 \\ \lambda_3 = aA^{3/2} - 27\,a^4 + 18\,a^2A - 72\,a^3 + 144\,a^2b + 8\,aA - 64\,bA - 56\,a^2 + 320\,ba \\ \qquad -128\,b^2 + 8\,A + 32\,a + 320\,b + 16 \end{cases}$$

where $A = \sqrt{9\,a^2 + 4\,a - 32\,b + 4}$. The equation of the genus 2 curve is

$$y^2 = (x-1)\prod_{i=1}^{4}(x-\alpha_i),$$

and elliptic curves have equations

$$E_1: y^2 = \prod_{i=1}^{3}(x-\lambda_i), \qquad E_2: y^2 = x\prod_{i=1}^{3}(x-\lambda_i).$$

Notice that we write the equation of genus 2 curve in terms of only 2 unknowns. We denote the Igusa invariants of C by J_2,J_4,J_6, and J_{10}. The absolute invariants of C are given in terms of these classical invariants:

$$i_1 = 144\frac{J_4}{J_2^2}, \quad i_2 = -1728\frac{J_2J_4 - 3J_6}{J_2^3}, \quad i_3 = 486\frac{J_{10}}{J_2^5}.$$

Two genus 2 curves with $J_2 \neq 0$ are isomorphic if and only if they have the same absolute invariants. Notice that these invariants of our genus 2 curve are polynomials in a and b. By using a computational symbolic package (as Maple), we eliminate a and b to determine the equation for the non-degenerate locus $\mathcal{L}_4$. The result is very long. We do not display it here.

7. FINAL REMARKS AND FUTURE WORK

Let $\mathcal{X}$ be a genus 2 curve defined over a field K, $charK=p\geq 0$, and $\mathbb{J}ac(\mathcal{X},\iota)$ its Jacobian, where ι is the principal polarization of $\mathbb{J}ac(\mathcal{X})$ attached to $\mathcal{X}$. Assume that $\mathbb{J}ac(\mathcal{X})$ is (n,n)- geometrically reducible with E_1 and E_2 its elliptic components.

In an upcoming project, we would like to study pairs of (E_1,E_2) elliptic components and try to determine their number (up to isomorphism over $\bar{k}$) when they are isogenous of degree N, for an integer $N\geq 2$. We denote by $\phi_N(x,y)$ the N-th modular polynomial. Two elliptic curves with j-invariants j_1 and j_2 are N-isogenous if and only if $\phi_N(j_1,j_2)=0$. The equation $\phi_N(x,y)=0$ is the canonical equation of the modular curve $X_0(N)$. The equations of $X_0(N)$ are well-known.

In [3], Beshaj et al. prove that there are only finitely many curves $\mathcal{X}$ (up to isomorphism) defined over K such that E_1 and E_2 are N-isogenous for $n=2$ and $N=2,3,5,7$ with $\mathrm{Aut}(\mathbb{J}ac\mathcal{X}) \cong V_4$ or $n=2$, $N=3,5,7$ with $\mathrm{Aut}(\mathbb{J}ac(\mathcal{X})) \cong D_4$. The same holds if $n=3$ and $N=5$. Furthermore, by determining the Kummer and the Shioda-Inose surfaces for the above $\mathbb{J}ac(\mathcal{X})$ we can show how such results in positive characteristic $p>2$ suggest nice applications in cryptography. Now that we have computed the locus $\mathcal{L}_4$, it would be interesting to explore the same problem when $n=4$ and $N=2,3,5,7$.

ACKNOWLEDGMENTS

The authors would like to thank our supportive colleagues at the Army Cyber Institute at West Point.

REFERENCES

[1] M. Ajtai, "*Generating hard instances of lattice problems (extended abstract)*," In Proc. Twenty-eighth Annual ACM Symposium on the Theory of Computing, Philadelphia, PA, 1996, ACM, New York, 1996, pp. 99–108. MR 1427503.

[2] J. Alwen, *What is lattice-based cryptography and why should you care*. Publication details? 2018.

[3] L. Beshaj, A. Elezi and T. Shaska, "Isogenous components of jacobian surfaces," *European Journal of Mathematics*, 2019.

[4] Charles, D., & Lauter, K. (2005). Computing Modular Polynomials. LMS Journal of Computation and Mathematics, 8, 195-204. doi:10.1112/S1461157000000954.

[5] L. Chen, S. Jordan, Y. Liu, D. Moody, R. Peralta and D. Smith-Tone, "Report on post-quantum cryptography," National Institute of Standards and Technology Internal Report, NIST.IR.8105. 2016.

[6] S. Contini, A. K. Lenstra and Ron Steinfeld, "VSH, an efficient and provable collision-resistant hash function," Advances in cryptology – EUROCRYPT 2006, Lecture Notes in Comput. Sci., vol. 4004, Springer, Berlin, 2006, pp. 165–182. MR 2423542.

[7] C. Costello and H Hisil, "A simple and compact algorithm for sidh with arbitrary degree isogenies," presented at International Conference on the Theory and Application of Cryptology and Information Security ASIACRYPT, Advances in Cryptology ASIACRYPT (2017), 303–329.

[8] Costello C. (2018) Computing Supersingular Isogenies on Kummer Surfaces. In: Peyrin T., Galbraith S. (eds) Advances in Cryptology – ASIACRYPT 2018. ASIACRYPT 2018. Lecture Notes in Computer Science, vol 11274. Springer, Cham.
[9] L. De Feo, D. Jao and J. Plút, "Towards quantum-resistant cryptosystems from supersingular elliptic curve isogenies," *J. Math. Cryptol*., vol. 8, pp. 209–247, 2014.
[10] L. Ducas, A. Durmus, T. Lepoint and V. Lyubashevsky, "*Lattice signatures and bimodal Gaussians*," Advances in cryptology – CRYPTO 2013. Part I, Lecture Notes in Comput. Sci., vol. 8042, Springer, Heidelberg, 2013, pp. 40–56. MR 3126416.
[11] L. Ducas and D. Micciancio, "Improved short lattice signatures in the standard model," Advances in cryptology – CRYPTO 2014. Part I, Lecture Notes in Comput. Sci., vol. 8616, Springer, Heidelberg, 2014, pp. 335–352. MR 3239444.
[12] L. Feo, *Mathematics of isogeny based cryptography*, Arxiv, 2017.
[13] G. Frey and T. Shaska, "Curves, Jacobians, and Cryptography, Algebraic curves and their applications" (L. Beshaj, ed.), Contemporary Math., vol. 724, American Mathematical Society, 2019, pp. 280–350.
[14] K. Hashimoto and N. Murabayashi, "Shimura curves as intersections of Humbert surfaces and defining equations of QM-curves of genus two," *Tohoku Math*. J., vol. 2, no. 47, pp. 271–296, 1995. MR 1329525.
[15] A. Hülsing, J. Rijneveld and F. Song, *Mitigating multi-target attacks in hash-based signatures*, Public-key cryptography – PKC 2016. Part I, Lecture Notes in Comput. Sci., vol. 9614, Springer, [Cham], 2016, pp. 387–416. MR 3492589.
[16] J. Igusa, "*Arithmetic variety of moduli for genus two*," Ann. of Math., vol. 2, no. 72, pp. 612–649, 1960. MR 0114819.
[17] D. Jao and L. De Feo, "Towards quantum-resistant cryptosystems from supersingular elliptic curve isogenies," Post-quantum cryptography, Lecture Notes in Comput. Sci., vol. 7071, Springer, Heidelberg, 2011, pp. 19–34. MR 2931459.
[18] N. Koblitz, "Elliptic curve cryptosystems," *Math. Comp*., vol. 48, no. 177, pp. 203–209, 1987. MR 866109.
[19] A. Kumar, "Hilbert modular surfaces for square discriminants and elliptic subfields of genus 2 function fields," *Res. Math. Sci*., vol. 2, Art. 24, 46, 2015. MR 3427148.
[20] D. Lombardo, *Computing the geometric endomorphism ring of a genus 2 jacobian* Math. Comp. 88 (2019), 889-929.
[21] V. Lyubashevsky, "Lattice signatures without trapdoors," Advances in cryptology – EUROCRYPT 2012, Lecture Notes in Comput. Sci., vol. 7237, Springer, Heidelberg, 2012, pp. 738–755. MR 2972929.
[22] K. Magaard, T. Shaska, and H. Völklein, "*Genus 2 curves that admit a degree 5 map to an elliptic curve*," *Forum Math*., vol. 21, no. 3, pp. 547–566, 2009. MR 2526800.
[23] A. Malmendier and T. Shaska, A universal genus-two curve from Siegel modular forms, SIGMA. Symmetry, Integrability and Geometry. Methods and Applications, vol 13 (2017), no. 089, 17 pages. MR 3731039.
[24] A. Malmendier and T. Shaska, "From hyperelliptic to superelliptic curves," *Albanian J. Math*., vol. 13, no. 1, pp. 107–200, 2019. MR 3978315.
[25] R. J. McEliece, A Public-key cryptosystem based on algebraic coding theory, DSN Progress Report, Jet Propulsion Laboratory, Pasadena, CA (Jan./Feb. 1978) pp. 114–116.
[26] V. S. Miller, "Use of elliptic curves in cryptography," Advances in cryptology – CRYPTO '85 (Santa Barbara, Calif., 1985), Lecture Notes in Comput. Sci., vol. 218, Springer, Berlin, 1986, pp. 417–426.
[27] H. Niederreiter, "*Knapsack-type cryptosystems and algebraic coding theory," Problems Control Inform. Theory/Problemy Upravlen. Teor. Inform*., vol. 15, no. 2, pp. 159–166, 1986. MR 851173.
[28] N. Pjerro, M. Ramosaco and T. Shaska, "Degree even covering of elliptic curves by genus 2 curves," *Albanian Journal of Mathematics*, vol. 2, no. 3, pp. 241–248, 2008.
[29] O. Regev, "On lattices, learning with errors, random linear codes, and cryptography," J. ACM, vol. 56, no. 6, art. 34, p. 40, 2009. MR 2572935.
[30] F. Richelot, "Essai sur une methode generale pour determiner la valuer des integrales ultra-elliptiques, fondee sur des transformations remarquables des ce transcendantes," *CR Acad. Sc. Paris*, vol. 2, pp. 622–627, 1836.
[31] T. Shaska, "Curves of genus 2 with (N,N) decomposable Jacobians," *J. Symbolic Comput*., vol. 31, no. 5, pp. 603–617, 2001. MR 1828706.
[32] T. Shaska, "Curves of genus two covering elliptic curves," ProQuest LLC, Ann Arbor, MI, Thesis (Ph.D.), University of Florida, 2001.
[33] T. Shaska, "Genus 2 curves with (3,3)-split Jacobian and large automorphism group," *Algorithmic number theory*, Lecture Notes in Comput. Sci., vol. 2369, Springer, Berlin, pp. 205–218, 2002.

[34] T. Shaska, “Genus 2 fields with degree 3 elliptic subfields,” *Forum Math*., vol. 16, no. 2, pp. 263–280, 2004. MR 2039100.
[35] T. Shaska and H. Völklein, “Elliptic subfields and automorphisms of genus 2 function fields,” *Algebra, arithmetic and geometry with applications*, Springer, Berlin, pp. 703–723, 2004. MR 2037120.
[36] T. Shaska, G. S. Wijesiri, S. Wolf, and L. Woodland, “Degree 4 covering of elliptic curves by genus 2 curves,” *Albanian Journal of Mathematics*, vol. 2, no. 4, pp. 307–318, 2008.
[37] Peter W. Shor, “Algorithms for quantum computation: discrete logarithms and factoring” 35th Annual Symposium on Foundations of Computer Science (Santa Fe, NM, 1994), IEEE Comput. Soc. Press, Los Alamitos, CA, 1994, pp. 124–134. MR 1489242.
[38] J. Silverman and J. Tate, Rational points on elliptic curves, ISBN: 978-3-319-18587-3, Number of Pages : XXII, 332, 2nd edition, 2015.
[39] Stehlé D., Steinfeld R., Tanaka K., Xagawa K. (2009) Efficient Public Key Encryption Based on Ideal Lattices. In: Matsui M. (eds) Advances in Cryptology – ASIACRYPT 2009. ASIACRYPT 2009. Lecture Notes in Computer Science, vol 5912, pg 617-635. Springer, Berlin, Heidelberg.
[40] L. C. Washington, *Elliptic Curves: Number Theory and Cryptography*, second edition, London: Chapman and Hall/CRC, 2008.
[42] M. Zhandry, “A note on the quantum collision and set equality problems,” *Quantum Inf. Comput*., vol. 15, no. 7-8, pp. 557–567, 2015.

BIOGRAPHIES

Editors

Taťána Jančárková is a researcher in the Law Branch of NATO CCDCOE where she works on the application of international law to cyberspace operations (the Cyber Law Toolkit project) and cybersecurity-related aspects of 5G technologies. She has previously served as legal adviser and head of unit at the National Cyber and Information Security Agency of the Czech Republic, her responsibilities including negotiations of EU cyber legislation and cyber defence related cooperation with NATO and OSCE. She holds degrees in Law and in Russian and East European Studies from Charles University in Prague and an LL.M. in Public International Law from Leiden University.

Lauri Lindström has been a researcher at NATO CCDCOE since May 2013. Prior to that, he worked in the Estonian Ministry of Foreign Affairs (2007–2012) as Director General of Policy Planning and held various positions in the Ministry of Defence (1995–2007), dealing mainly with issues related to international cooperation, Estonia's accession to NATO, defence planning and security policy. Lauri holds a PhD from Tallinn University, Estonia.

Massimiliano Signoretti is a Lieutenant Colonel in the Italian Air Force and a researcher in the Law Branch of NATO CCDCOE. His research area is public international law, international humanitarian law and the law of armed conflict. He graduated in law at the University of Rome (La Sapienza) and was admitted to the Bar. He also studied at the University of Stockholm Faculty of Law and has worked at the Italian Defence General Staff – Office of International Legal Affairs. His career also includes four years' service at the NATO Partnerships Division, ACO SHAPE, Belgium. His academic achievements include a Master's degree in Strategic International Military Studies and a level II Master's degree in International Humanitarian Law and the Law of Armed Conflict.

Lt. **Ihsan Tolga** is a researcher at NATO CCDCOE. Prior to taking up this post, he worked for three years in the Turkish Navy Research Centre Command and before that as an officer in the Turkish Navy Submarine Fleet Command. He is a graduate of the Turkish Naval Academy (BSc in Computer Engineering) and the University of Southern California (MSc in Computer Science).

Maj. **Gábor Visky** is a researcher in the Technology Branch of NATO CCDCOE, where his main field of expertise is industrial control systems. Gábor's previous assignments include 15 years of designing hardware and software for embedded

control systems and researching their vulnerabilities through reverse engineering. Gábor holds an MSc degree in Information Engineering with a specialty in Industrial Measurement and a BSc in the field of telecommunication.

Co-Editors

Lt.Col. **Henrik Paludan Beckvard** has both an army and legal background and has served in various staff positions in Defence Command Denmark, the Danish Home Guard Command and the Ministry of Defence. Since 2018, Henrik Beckvard has served as a researcher in the Strategy Branch of NATO CCDCOE in Tallinn, Estonia. Among his tasks, Henrik is on the team preparing the strategic track for the annual exercise Locked Shields and also serves as the CCDCOE Course Director for Critical Information Infrastructure Protection. Together with Lauri Lindström, Henrik also conducts the flag officer level Executive Cyber Seminar.

Capt. **Costel-Marius Gheorghevici** is a researcher in the Technology Branch of NATO CCDCOE. His research focus is on the cybersecurity of wireless networks and telecommunications infrastructure. Prior to taking up this post, he worked as an electronic warfare officer in the Romanian MoD. He holds an MSc degree in Electronics and Telecommunication Engineering.

Kadri Kaska is Head of the Law Branch at NATO CCDCOE and was formerly a cyber security policy and legal researcher at the same organisation. Her research has revolved around the issues of national cybersecurity strategy and governance and the legal aspects of state cyber activities. In 2017–2018, she served as a lead analyst at the Estonian Information System Authority, contributing to the Agency's activities in cyber threat assessment, policy analysis and legal drafting. She was the lead author and editor of the annual Estonian Cyber Security Assessments and one of the authors of Estonia's new Cyber Security Act and the 2018 national cybersecurity strategy.

Liina Lumiste is an international law researcher at NATO CCDCOE. Prior to joining CCDCOE, she worked as legal counsel for the Ministry of Education and Research. Since 2018, Liina has also been an assistant lecturer at the University of Tartu. Her research areas include public international law, the legislative processes of international law, non-state actors in international law, international humanitarian law and international criminal law. She has a Master's degree in Law from the University of Tartu.

Piret Pernik is a researcher in the Strategy Branch of NATO CCDCOE. Her main research areas are cyber security strategies and policies, horizon scanning and analysis of cyber threats, and the development of military cyber organisations. She has worked

as a researcher in cyber security since 2013 at the International Centre for Security and Defence and the Estonian Academy of Social Sciences. Her published work includes academic journal articles, think-tank policy analysis and research reports, and book chapters. She holds Master's degrees in Social Theory from the University of Tallinn, Estonian Institute of Humanities, and in International Relations and European Studies from the Central European University, Budapest.

Ann Väljataga works as a researcher at NATO CCDCOE, where her areas of expertise cover national cyber security strategies and public international law. She holds a BA in Law from the University of Tartu and a Master's degree in Law and Technology from Tallinn University of Technology. Previously, she conducted legal research at the Estonian Human Rights Centre, where she focused on privacy and data protection, and at the European Union Agency for Fundamental Rights (FRA), where her work examined the fundamental rights implications of untargeted surveillance and biometric border control systems.

Authors

Thibaut Alchus is an independent researcher specialising in Russian-speaking cyberspace. His research interests include hybrid threats, the isolation of the Russian Internet and the evolution of cyberspace in post-Soviet frozen conflicts. He holds a Master's degree in Comparative Military History from the Institute of Political Studies (Sciences Po Aix) and in Cyberstrategy & Data Science from the French Institute of Geopolitics. Thibaut is currently collaborating with the GEODE centre at the University of Paris 8.

Gilberto Pires de Azevedo holds a BSc degree in Electrical Engineering from PUC-Rio, an MSc from COPPE-UFRJ also in Electrical Engineering and a DSc in Computer Science from PUC-Rio – all in Rio de Janeiro, Brazil – and a specialisation in Strategic Management of Technological Innovation at UNICAMP, Campinas, Brazil. He has been a researcher at Cepel Electrical Energy Research Center in Rio de Janeiro since 1985, with experience in areas such as network analysis, human-machine interfaces, software development, multi-agent systems, EMS/SCADA, cybersecurity and R&D management.

Alastair R. Beresford is Professor of Computer Security and Deputy Head of the Department of Computer Science and Technology at the University of Cambridge. His research examines the security and privacy of large-scale distributed computer systems. Within this broad area, he is particularly interested in the security and privacy of networked mobile and embedded devices such as smartphones, tablets and laptops, as well as the Internet of Things and industrial control systems. His research examines

the security of the devices themselves as well as the security and privacy problems induced by the interaction between these devices and cloud-based Internet services. Previous work includes critical technical evaluation of existing products, designing and building novel prototype technologies, and measuring human behaviour.

Dr **Lubjana Beshaj** is a Cyber Fellow of Mathematics at the Army Cyber Institute and an Assistant Professor in the Department of Mathematical Sciences at West Point. She has a BS in Mathematics and Physics from the University of Vlora (Albania), an MS in Mathematical Sciences from the University of Vlora (Albania), and a PhD in Applied Mathematics from Oakland University (USA). Her research interests include cryptography, elliptic and hyperelliptic curve cryptography, post-quantum cryptography (specifically isogeny-based cryptography), Jacobian varieties and the arithmetic of algebraic curves.

Maxli Barroso Campos is a Senior Officer of the Cyber Defense Command in Brazil and holds a Master's degree in Computer Systems from the University of Salvador (Brazil). He has CISSP, CEH and GISCP certification and has been working in the cyber area since 2011 in major events such as Rio +20, the Confederations Cup and World Cup, as well as in strategic projects of the Ministry of Defense. Currently, Maxli works as Head of the Security Division of the Command's Strategic Management Department and serves as one of the coordinators of the Cyber Guardian Exercise. In the academic arena, he is Adjunct Professor at the University Center of Brasília.

Joe Cheravitch has worked as an analyst focused on international cyber and information warfare since 2014. He served as a psychological operations specialist in the US Army from 2008 to 2012, deploying to Afghanistan in 2010 and Iraq in 2011. Joe Cheravitch holds an MS in Foreign Service from the Edmund A. Walsh School of Foreign Service at Georgetown University.

Michael Dodson is a PhD candidate at the Department of Computer Science and Technology at the University of Cambridge, where he researches usable and sustainable security for industrial systems.

Frédérick Douzet is Professor of Geopolitics at the University of Paris 8, director of the French Institute of Geopolitics research team (IFG Lab) and director of the Centre for Geopolitics of the Datasphere (GEODE). She is a Commissioner of the Global Commission on the Stability of Cyberspace (cyberstability.org) and is a member of the French Defense Ethics Committee since January 2020. In 2017, she was part of the drafting committee for the Strategic Review of Defense and National Security.

Keir Giles is a Senior Consulting Fellow with the Russia and Eurasia Programme at Chatham House in London and also serves as Research Director for the Conflict Studies Research Centre, formerly part of the UK Ministry of Defence. Keir has been involved with issues surrounding the exploitation of the Internet for three decades and combines a technical background with in-depth study of authoritarian regimes' approaches to information security in order to develop analysis and prediction of the development of information warfare, including the subdomain of cyber conflict. He is the author of ground-breaking studies on Russian theory, doctrine and structures for engaging in information and cyber confrontation.

Roman Graf, PhD, OSCP, CEH, research engineer at the Center for Digital Safety & Security in the Austrian Institute of Technology GmbH, works on cyber security and data analytics topics, contributing to the development of several European research projects like Titanium, MAL2, Ecossian, E-ARK, DMA, EDSI, Planets, Assets and SCAPE. He has published widely in the area of cyber security and risk management in digital preservation, being an active member of the Open Preservation Foundation (OPF). Dr Graf supported the development of cyber threat intelligence solution CAESAIR, serving as one of the key developers, and he contributed a module to the Open Source Threat Intelligence Platform (MISP).

Col. **Andrew O. Hall** is the Director of the Army Cyber Institute at the United States Military Academy located at West Point, New York, and he serves as the Chairman of the Editorial Board for *The Cyber Defense Review* journal. He has a BS in Computer Science from USMA, an MS in Applied Mathematics from the Naval Postgraduate School, and a PhD in Management Science from the University of Maryland. Colonel Hall's military career has been focused on operations research and solving the army's most challenging problems using advanced analytic methods. His research interests include data science, cyber education and applied probability.

Jakub Harašta is an Assistant Professor at the Institute of Law and Technology, Masaryk University, Brno. Jakub holds a Master's degree in Law (2013), a PhD in Law of Information and Communication Technologies (2018), and a Master's degree in Security and Strategic Studies (2020). Jakub was a postdoctoral fellow (non-resident) at the Center for Cyber Law and Policy, Haifa University (2018). He currently holds the position of editor-in-chief of Masaryk University's *Journal of Law and Technology*. In his research, Jakub focuses on legal informatics and cybersecurity.

Kim Hartmann is the Cyber and Information Technology Director at the Conflict Studies Research Centre, formerly part of the UK Ministry of Defence. She is a senior consultant and researcher and an acknowledged evaluator for EU projects in the fields of cyber, network and software security. As a computer scientist and mathematician,

she combines profound technical knowledge with an in-depth analysis of the geopolitical context of cyber incidents. Her fields of excellence are cyber security risk-assessment of embedded systems and networks, IT forensics, privacy protection and secure software development. She has advised private, military and governmental organisations for more than a decade on the assessment and integration of security aspects in existing technologies.

Jason Healey is Senior Research Scholar at Columbia University's School for International and Public Affairs, specialising in cyber conflict and risk. He started his career as a US Air Force intelligence officer before moving to cyber response and policy jobs at the White House and Goldman Sachs. He was the founding Director for Cyber Issues at the Atlantic Council, where he remains a Senior Fellow and is the editor of the first history of conflict in cyberspace, *A Fierce Domain: Cyber Conflict, 1986 to 2012*. He is on the DEFCON review board and served on the Defense Science Board task force on cyber deterrence.

Kimmo Heinäaro is a researcher in the Technology Branch of NATO CCDCOE. Prior to this current assignment, he worked as a cyber researcher in the Finnish Defence Research Agency. He is currently researching the cyber security of ICS, SCADA and embedded systems.

Neil Jenkins is a Chief Analytic Officer at the Cyber Threat Alliance. Neil leads the CTA's analytic efforts, focusing on the development of threat profiles, adversary playbooks and other analysis using the threat intelligence in the CTA platform. Previously, he served in various roles within the U.S. Department of Homeland Security, Department of Defense, and Center for Naval Analyses, where he spearheaded numerous initiatives tied to cybersecurity strategy, policy and operational planning for both the public and private sectors.

Aleksi Kajander is an MA candidate at the Tallinn University of Technology for Law and Technology. Alongside his studies, he practises law at a business law firm. He holds a Master's degree in Investment Treaty Arbitration from Uppsala University and a *cum laude* Bachelor's Law degree in EU and International Law. In 2016, he participated in the All-European International Humanitarian and Refugee Law Moot Court Competition as a part of the winning TalTech team and won the Best Oralist of the Final Round award. Prior to his studies, he served as a squad leader in the Finnish Rapid Deployment Forces.

Dr **Agnes Kasper** is a Senior Lecturer in Law and Technology in the Department of Law of the Tallinn University of Technology (TalTech). Dr Kasper holds diplomas in international business, law and management. She has received formal training

on technical aspects of cybersecurity and digital evidence. Dr Kasper has served at embassies and human rights organisations and she led the legal department in an IT consultancy and development company. She has also acted in an advisory capacity in consultations with governments on issues relating to cybersecurity. Her research focuses on regulatory aspects of cybersecurity.

Jeff Kosseff is an Assistant Professor of Cybersecurity Law in the United States Naval Academy's Cyber Science Department. His latest book, *The Twenty-Six Words That Created the Internet, a History of Section 230 of the Communications Decency Act*, was published in Spring 2019 by Cornell University Press. He is also the author of *Cybersecurity Law*, a textbook and treatise published by Wiley in 2017, with a second edition released in November 2019. Jeff practised cybersecurity, privacy and First Amendment law at Covington & Burling, and clerked for Judge Milan D. Smith Jr. of the United States Court of Appeals for the Ninth Circuit and Judge Leonie M. Brinkema of the United States District Court for the Eastern District of Virginia. Before becoming a lawyer, he was a technology and political journalist for *The Oregonian* and was a finalist for the Pulitzer Prize for national reporting and recipient of the George Polk Award for national reporting. He received a JD from Georgetown University Law Center and a BA and MPP from the University of Michigan.

Ivana Kudláčková is a research fellow at the Institute of Law and Technology, Faculty of Law, Masaryk University (the Czech Republic). She is a member of the research team of the project CyberSecurity, CyberCrime and Critical Information Infrastructure Centre of Excellence. She predominantly focuses on the use of force in cyberspace, new methods of warfare and the relevance of international public law in cyberspace. She graduated in law from Masaryk University and studied further at the Georgian Institute of Public Affairs (Georgia) and at Ghent University (Belgium).

Neal Kushwaha is the recipient of the 2019 Royal Humane Silver Medal of Bravery award, an international multilingual and motivational speaker, and guest lecturer at Stellenbosch University. He is the founder of a Canadian consulting company, IMPENDO Inc., which specifically serves public-sector clients. His research and consulting fall within the cross-section of cyberspace, security and law. Neal has publications at various venues on topics ranging from policy and doctrine to technical matters. Aligned with his PhD research, he supports nations with their cyber programmes. While he is a prominent speaker at cyber conferences on the global stage, it is his climbing that generates the most interest: he is an accomplished mountaineer with more than 27 years of climbing experience and he leads climbs, including Everest, under the banner of *Big Climbs*.

Artūrs Lavrenovs is a researcher at NATO CCDCOE focusing on the web and network technologies while teaching security courses, performing applied and academic research, and contributing to cyber exercises. Arturs has taught web technology and IT security courses at the University of Latvia, where he is currently a PhD student doing research in the cyber security domain.

Martin C. Libicki (PhD, UC Berkeley 1978) holds the Keyser Chair of Cybersecurity Studies at the US Naval Academy. In addition to teaching, he carries out research in cyberwar and the general impact of information technology on domestic and national security. He is the author of a 2016 textbook on cyberwar, *Cyberspace in Peace and War*, as well as *Conquest in Cyberspace: National Security and Information Warfare* and various related RAND monographs. Prior employment includes 12 years at the National Defense University, three years on the Navy Staff (Logistics) and three years for the US GAO.

Bilyana Lilly is a Pardee Fellow at the Pardee RAND Graduate School of the RAND Corporation. She leads project teams and co-authors reports on Russian cyber threat actors, information warfare, election cyber security, disinformation, machine learning for text analysis and NATO cybersecurity. She has presented RAND-sponsored and independent research at professional conferences, including CyCon, DefCon and the Warsaw Security Forum. Prior to joining RAND, Lilly was an associate at the Brookings Institution, where she focused on US security strategy and NATO's policy toward emerging powers. She is the author of the book *Russian Foreign Policy Toward Missile Defense*, which contains information from talks that Lilly conducted with Russia's Deputy Defence Minister and Deputy Minister of Foreign Affairs. Lilly has Master's degrees from Oxford University (distinction), UK, and the Graduate Institute of International and Development Studies, Switzerland.

Kevin Limonier is an Associate Professor in Geography and Slavic Studies at the French Institute of Geopolitics (University of Paris 8), and is Vice Director of GEODE (www.geode.science), a research centre dedicated to the geopolitics of the datasphere. He is head of the Observatory of Russian-Speaking Cyberspace, a research unit dedicated to the post-Soviet segment of digital space as a new geopolitical and technical object of studies. His work focuses on the history and geography of Russian cyberspace and on the development of new methods of data collection and data visualisation for mapping digital exchanges, borders and conflicts in Eurasia. He is also a specialist in innovation policies in the USSR and contemporary Russia.

Livinus Nweke is a PhD candidate with the Department of Information Security and Communication Technology (IIK) at the Norwegian University of Science and Technology (NTNU) Gjøvik, Norway. Prior to joining NTNU, Livinus received his

MSc degree in Computer Science (*summa cum laude*) from the Faculty of Ingegneria dell'Informazione, Informatica e Statistica, Sapienza Universita di Roma, Italy. He also worked briefly as a customer support engineer with the Next-Generation Firewall Team at Cisco Inc. Kraków, Poland. His research interests include, but are not limited to, critical infrastructure protection, software-defined networking (SDN) and general information security.

Dr **Tina J. Park** is a co-founder and Executive Director of the Canadian Centre for the Responsibility to Protect based at the University of Toronto. She is also a Vice-President of the NATO Association of Canada and a former Fellow at the NATO Defense College in Rome. Dr Park is a globally renowned expert on North Korea's nuclear weapons programme and has advised over 30 governments and international organisations on their human rights policies, including the UN Secretary-General Ban Ki-Moon on R2P. She is a frequent media commentator for CBC, BBC, CTV, CBS and Al Jazeera. (www.tinapark.ca) @JIWONTINA

Paulo César Pellanda holds a BSc degree from the Federal University of Technology – Paraná (Brazil) and an MSc degree from the Military Institute of Engineering – IME (Brazil), both in Electrical Engineering; he also holds a PhD in Automatic Control from the French National Higher Institute of Aeronautics and Space (ISAE, Toulouse), where he received the Best Thesis Prize in Automatic Control from the French National Center for Scientific Research. With over 35 years' experience as a research scientist and professor, his main research projects have included the development of optimal robust control algorithms applied to multivariable dynamical systems in aerospace and power system fields. He is currently an Associate Professor at IME.

Louis Pétiniaud is a PhD student at the French Institute of Geopolitics (University of Paris 8) and researcher at the Geopolitics of the Datasphere (GEODE) centre. His research focuses on the geopolitics of the Black Sea. His main area of study considers the growing importance of the Internet's physical infrastructures and BGP routing topology in disputed territories in Georgia and Ukraine. He is also working on developing methods for mapping and visualising the transit of digital data. He benefits from the support of young researchers at the Institut des Hautes Etudes de Défense Nationale.

Dr **Przemysław Roguski** is a Lecturer in Law at the Jagiellonian University in Kraków (Poland) and an expert on cybersecurity and international law at the Kościuszko Institute. His research focuses on the law of peacetime cyber operations and different aspects of international law relating to cybersecurity, ICT and Internet governance. Previously, Przemysław has worked in private practice and as a lecturer for the German Academic Exchange Service (DAAD). He holds law degrees from the

University of Mainz (Germany) and Trinity College Dublin (Ireland) and a PhD in International Law from Jagiellonian University.

Kavé Salamatian is a full Professor of Computer Science at the University of Savoie. His main area of research has been Internet measurement and modelling, network security and networking information theory. He was previously a reader at Lancaster University, UK, and an Associate Professor at the University Pierre et Marie Curie. Kavé graduated in 1998 from Paris Sud-Orsay University, where he worked on joint source channel coding applied to multimedia transmission over the Internet for his PhD. In a former life, he graduated with an MBA and worked on the market floor as a risk analyst and enjoyed being an urban traffic modeller for some years. He is currently a Distinguished Visiting Professor at the Chinese Academy of Science and also works closely with the GEODE research group on geopolitics. He was the recipient of the Chinese Academy of Science Presidential Award in 2018.

Loqman Salamatian is an associate researcher at the GEODE centre. His main interests are Internet measurements, complex systems analysis and applied mathematics. His work includes mean-field analysis, computational geometry and Ricci flow applications for assessing stability in complex networks. He is focusing on the transmission protocols of the Internet, specifically on finding new ways of measuring interactions between virtual and geographical spaces.

Dr **Matthias Schulze** is a researcher at the security division of the German Institute for International and Security Affairs – Stiftung Wissenschaft & Politik (SWP). His research focuses on the dark side of digitisation, namely, the strategic use of cyber capabilities in international relations, cyber conflicts, cyber espionage, information operations and cybercrime. He also works on numerous domestic policy-related topics, such as encryption, vulnerability disclosure, government hacking and lawful access. Matthias holds a PhD in International Relations from the Friedrich-Schiller-University in Jena, Germany.

Christoph Steup received a Master of Science in Technology (Electrical Engineering) degree from Helsinki University of Technology.

Michael Switzer is a graduate of Trinity College at the University of Toronto and a JD candidate at the Allard School of Law at the University of British Columbia. He has served as the Deputy Executive Director at the Canadian Centre for the Responsibility to Protect since 2018 and led the Centre's delegation to the UN General Assembly's Dialogue on R2P at the UNHQ in New York in 2016 and 2017.

Evhen Tsybulenko is an Estonian legal scholar of Ukrainian descent. In 2000, he earned a PhD in International Law from Kiev National University, Ukraine, and carried out postdoctoral research at the International Human Rights Law Institute of De Paul University, USA in 2002. In Ukraine, he worked at the International Committee of the Red Cross (ICRC) and Kiev International University, where he has been a Professor since 2009. In Estonia, he was elected as Professor of Law (2005) and appointed Chair of the International and Comparative Law Department at the Tallinn University of Technology's Law School (2005–2010). Currently he is a Senior Lecturer at the same school. He was a founder and Director of the Tallinn Law School's Human Rights Centre (2007–2014). He was also an Adjunct (Visiting) Professor and Senior Visiting Mentor at the Joint Command and General Staff Course (JCGSC) at the Baltic Defence College (2007–2013). He has published more than 40 books and academic articles and more than 200 general interest articles, comments and interviews in 15 countries, mainly in Ukrainian or Russian, but also in English and other languages. He cooperates, as an external expert, with the ICRC, Estonian Red Cross, Estonian Integration Commission, Directorate-General for Education and Culture of the European Commission and the Organization for Security and Co-operation in Europe (OSCE).

Mikael Vingaard currently works in a government role and is one of the leading experts in the realm of deception technology (aka honeypots) in industrial environments. He has more than 20 years of experience in IT/OT Security (Red team/Blue team/Purple team) and enjoys doing active defence/threat hunting within industrial control systems (ICS). As part of building a safer tomorrow, Mikael has been credited for finding (and responsibly disclosing) 15+ Zero-days to leading IT and industrial vendors.

David Wallace is a Colonel in the United States Army. He is Professor and Head of the Department of Law, United States Military Academy, West Point, New York, where he has served on the faculty for 19 years. He has also served as a Deputy Staff Judge Advocate; Assistant/Associate Professor at the Judge Advocate General's School of the Army; Trial Attorney, Contract Appeals Division, United States Army Legal Service Agency; Trial Counsel and Legal Assistance Attorney, 3rd Infantry Division; and Public/Civil Affairs Officer, 81st Infantry Brigade. Colonel Wallace deployed to Afghanistan in 2004 and 2010 in support of Operation Enduring Freedom. In 2004, he served as a member of an implementation team working to establish the National Military Academy of Afghanistan (NMAA). He is a professor of discipline with an expertise in International Humanitarian Law. He also served as a visiting scholar at NATO CCDCOE in 2017.

Prof **Bruce Watson** (PhD) is Chief Scientist at IP Blox, where he provides pattern recognition, algorithmic and AI solutions for cybersecurity. Additionally, he consults

to public-sector organisations on AI and cybersecurity, while being involved in some start-ups and serving as full Professor at Stellenbosch University (Centre for Artificial Intelligence Research). Watson's first PhD was in Computing Science and Engineering from Eindhoven, after studying discrete mathematics and computer science in Waterloo (Canada). He later returned to Eindhoven as chair of Software Construction. Watson's second PhD is from the University of Pretoria. Parallel to his academic career, he worked as a compiler engineer at several companies (e.g. Microsoft), followed by engineering and architecture work on pattern recognition for security (e.g. for Cisco). Watson has been a presenter, participant and reviewer at several CyCons, mostly recently moderating the Quantum Computing panel in 2018 and the Artificial Intelligence panel in 2019.

Dr **Christopher Whyte** is Assistant Professor at Virginia Commonwealth University in the L. Douglas Wilder School of Government and Public Affairs Program on Homeland Security and Emergency Preparedness. He teaches coursework on cybersecurity policy, cyber conflict dynamics, political risk analysis and strategic planning. His research interests include a range of international security topics related to the use of information technology in war and peace, political communication and cybersecurity doctrine/policy. His published work, which includes two books and numerous articles, examines the contours of digital age information warfare, decision-making in cyber operations, online fringe social spaces and artificial intelligence.

Stephen Wolthusen holds dual appointments as Professor of Information Security at the Department of Information Security and Communication Technology at the Norwegian University of Science and Technology (NTNU) and as Professor of Information Security at the Department of Information Security at Royal Holloway, University of London. He has published more than 200 peer-reviewed articles and authored or edited several books with a focus on the modelling and analysis of cyber-physical systems security and network security. He has led and participated in a number of national and international research projects in these areas and serves on national and European advisory bodies.

JD Work serves as the Bren Chair for Cyber Conflict and Security at Marine Corps University and as a non-resident Senior Fellow with the Atlantic Council's Cyber Statecraft Initiative. He holds additional affiliations with the School of International and Public Affairs at Columbia University and the Elliot School of International Affairs at George Washington University, and is a senior adviser to the Cyberspace Solarium Commission. He can be found on Twitter @HostileSpectrum.