2021

13th International Conference on Cyber Conflict: Going Viral

T. Jančárková, L. Lindström,
G. Visky, P. Zotz (Eds.)

CYCON

IEEE
Advancing Technology
for Humanity

2021
13TH INTERNATIONAL CONFERENCE ON CYBER CONFLICT: GOING VIRAL

COPYRIGHT AND REPRINT PERMISSIONS

NATO COOPERATIVE CYBER DEFENCE CENTRE OF EXCELLENCE

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

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

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

CYCON 2021 SPONSORS

TABLE OF CONTENTS

INTRODUCTION

When preparing the CyCon 2020 proceedings last year, the editors were confident that by the time of CyCon 2021, the COVID-19 pandemic would be over and Estonia's capital Tallinn would again welcome cyber enthusiasts from the four corners of the world to discuss cyber defence and security through the lens of policy, strategy, law and technology.

Developments have proven us wrong, however, and we continue learning how to work and meaningfully exchange ideas in the virtual world. In that vein, CyCon 2021 – Going Viral – has also gone virtual. This year's central theme alludes not only to the immediate implications of human crises for cyberspace; it also sets out to encourage discussion on the impact of the rapid proliferation and high unpredictability that processes in cyberspace are prone to, and the real-life implications these phenomena have. We need to acknowledge these, study them and strive to use them for our common benefit.

To our satisfaction, CyCon authors were not intimidated by the circumstances and have responded richly to the call for papers. It would also seem that the extraordinary times have inspired a good deal of unconventional thinking about cyberspace. Some papers even look boldly into the distant cyber future. We all know, however, that what may sound far-fetched today, may become a reality within a generation.

As usual, articles in this book reflect the three tracks of CyCon. Of a total of 20 articles, there are four legal, six technical and ten strategy papers.

On the legal track, the discussion revolves around norms of behaviour in cyberspace and innovative applications of principles of international humanitarian law to cyber operations. **François Delerue** opens with the queen of international law rules and principles and calls upon States to be bolder on sovereignty in cyberspace. **Vladimir Radunović, Jonas Grätz-Hoffmann** and **Marilia Maciel** add a private sector perspective to the implementation of norms in cyberspace. **Anastasia Roberts** and **Adrian Venables** then attempt to alleviate legal concerns stemming from the use of artificial intelligence in the targeting process. **Monica Kaminska, Dennis Broeders** and **Fabio Cristiano** take more of a policy approach and conclude the legal bloc by examining whether principles of distinction, precaution and discrimination could inspire a new norm regulating under-the-threshold cyber operations.

This year, policy considerations are not foreign to technical papers either, and vice versa, technical aspects serve as a springboard for conclusions on the strategy track.

Several papers explore cyber threats in the context of a specific industry or category of services. **Johannes Klick, Robert Koch** and **Thomas Brandstetter** provide a topical and practical study of the attack surface of the German healthcare sector. **Bobby Vedral** examines the vulnerability of the financial system to a systemic cyber attack. **Charles Harry** and **Skanda Vivek**, in their turn, look into cyber threat implications for the US commercial air sector, while **Keir Giles** and **Kim Hartmann** focus on and explore the critical dependencies in the communications sector.

James Pavur, Martin Strohmeier, Vincent Lenders and **Ivan Martinovic** sound the alarm with regard to technologies used to launch space missions and the policy implications of their vulnerabilities. **Csaba Krasznay** and **Gergő Gyebnár** offer a case study illustrating the challenges of cyber threat intelligence sharing in the energy sector.

Of course, pure policy papers cannot be absent from this year's selection. **Brandon Valeriano** and **Benjamin Jensen** present a unique insight into the work of the US Cyberspace Solarium Commission and identify the lessons learned. **Nikola Pijović** evokes the 19th-century Great Game and examines how modern powers compete to shape global cyberspace norms. **Jason Blessing** documents the growth in institutionalised cyber capabilities across the globe, thus helping us realise the evolving paradigm in states' cyber defence policies in and beyond NATO. **James Shires** takes an innovative look at disinformation operations and introduces their stratification in order to better understand their policy implications. Looking well beyond the horizon is the paper by **Neal Kushwaha, Keir Giles, Tassilo Singer** and **Bruce Watson**, who propose a new regulatory concept of cyber personhood be considered for complex cyber systems of the future.

Before we reach the distant future, a closer study of existing or emerging technologies is appropriate. **Roman Graf** and **Artūrs Lavrenovs** follow up on their earlier work on using artificial intelligence (AI) to classify devices on the internet. A paper authored by **Roland Meier, Artūrs Lavrenovs, Kimmo Heinäaro, Luca Gambazzi** and **Vincent Lenders** contemplates the engagement of AI in cyber defence exercises. **Vasileios Mavroeidis, Ryan Hohimer, Tim Casey** and **Audun Jøsang** seek to demonstrate how commonly agreed-upon controlled vocabularies can be practically used to enrich cyber threat intelligence and infer new information at a higher contextual level. **Pietro Baroni, Federico Cerutti, Daniela Fogli, Massimiliano Giacomin, Francesco Gringoli, Giovanni Guida** and **Paul Sullivan** examine the interaction of AI and humans in cyber threat analysis. The book concludes with **Martin C. Libicki** and **David C. Gompert** offering policy recommendations on quantum communications as an instrument for better cyber security.

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 thank the reviewers, who have invested their time and expertise to help us make the final selection. We remain equally grateful to our authors and researchers, who have chosen CyCon over other platforms to present their original work. 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.

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 preparing this book and for their courage in navigating the unchartered waters of virtual conferencing. Our gratitude goes (in alphabetical order) to Liis Poolak and Jaanika Rannu of the CCDCOE Support Branch for logistics support, and to Henrik Beckvard, Sungbaek Cho, Marius Gheorghevici, Kadri Kaska, Piret Pernik, Massimiliano Signoretti, Ann Väljataga and Jan Wünsche for their invaluable editorial assistance.

CyCon 2021 Programme Committee:

- Lauri Lindström, chair
- Taťána Jančárková, co-chair, chief editor of the proceedings
- Maj. Gábor Visky, co-chair
- Philippe Zotz, co-chair
- Maj. Vasileios Anastopoulos
- Lt. Col. Henrik Paludan Beckvard
- Capt. Costel-Marius Gheorghevici
- Maj. Emre Halisdemir
- Liina Lumiste
- Lt. Col. Gry-Mona Nordli
- Piret Pernik
- Lt. Col. Dr Massimiliano Signoretti
- Jan Wünsche

Academic Review Committee Members for CyCon 2021

- Siim Alatalu, Information System Authority, Estonia
- Maj. Geert Alberghs, Ministry of Defence, Belgium
- Maj. Vasileios Anastopoulos, NATO CCDCOE
- Daniel Peder Bagge, NÚKIB, Czech Republic

- Lt. Col. Henrik Paludan Beckvard, NATO CCDCOE
- Jacopo Bellasio, RAND Europe, Belgium
- Dr Bernhards Blumbergs, CERT.LV, Latvia
- Lt. Col. Pascal Brangetto, Ministère des Armées, France
- Dr Ben Buchanan, Georgetown University, United States
- Dr Joe Burton, University of Waikato, New Zealand Institute for Security and Crime Science, New Zealand
- Prof. Thomas Chen, City, University of London, United Kingdom
- Dr Sungbaek Cho, NATO CCDCOE
- Prof. Sean Costigan, George C. Marshall Center for European Security Studies, Germany
- Sebastian Cymutta, NATO CCDCOE
- Prof. Didier Danet, Military Academy of Saint-Cyr, France
- Samuele De Tomas Colatin, NATO CCDCOE
- Prof. Thibault Debatty, Royal Military Academy, Belgium
- Prof. Dorothy Denning, Naval Postgraduate School, United States
- Prof. Frédérick Douzet, GEODE, University Paris 8, France
- Dr Helen Eenmaa-Dimitrieva, University of Tartu, Estonia
- Amy Ertan, Royal Holloway, University of London, United Kingdom
- Dr Kenneth Geers, Atlantic Council, United States
- Capt. Costel-Marius Gheorghevici, NATO CCDCOE
- Keir Giles, Conflict Studies Research Centre, United Kingdom
- Prof. Michael Grimaila, Air Force Institute of Technology, United States
- Maj. Emre Halisdemir, NATO CCDCOE
- Dr Jonas Hallberg, Swedish Defence Research Agency, Sweden
- Dr Jakub Harašta, Masaryk University, Czech Republic
- Jason Healey, School of International and Public Affairs, Columbia University, United States
- Dr Trey Herr, Harvard University, United States
- Prof. David Hutchison, Lancaster University, United Kingdom
- Dr Ion Alexandru Iftimie, Cyber Security Cluster of Excellence, Romania
- Prof. Gabriel Jakobson, Altusys Corporation, United States
- Raik Jakschis, Hanse Digital Access OÜ, Estonia
- Taťána Jančárková, NATO CCDCOE
- Kadri Kaska, NATO CCDCOE
- Dr Károly Kassai, Cyber Defence Centre, Hungary
- Prof. Sokratis Katsikas, Open University of Cyprus, Cyprus
- Dr Panagiotis Kikiras, AGT R&D GmbH, Germany
- Dr Joonsoo Kim, National Security Research Institute, South Korea
- Dr Keiko Kono, NATO CCDCOE
- Prof. Csaba Krasznay, National University of Public Service, Hungary

- Capt. Juha Kukkola, Finnish National Defence University, Finland
- 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
- Dr Lauri Lindström, NATO CCDCOE
- Liina Lumiste, NATO CCDCOE
- Dr Kubo Mačák, International Committee of the Red Cross, Switzerland
- Youngjae Maeng, NATO CCDCOE
- Prof. Olaf Maennel, Tallinn University of Technology, Estonia
- Dr Matti Mantere, Luminor Bank, Estonia
- Prof. Aditya Mathur, Singapore University of Technology and Design, Singapore
- Dr Paul Maxwell, Army Cyber Institute, United States
- Markus Maybaum, Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie, Germany
- Dr Stefano Mele, Italian Atlantic Committee, Italy
- Tomáš Minárik, NÚKIB, Czech Republic
- Dr Jose Nazario, Fastly, United States
- Dr Lars Nicander, Swedish National Defence College, Sweden
- Lt. Col. Gry-Mona Nordli, NATO CCDCOE
- Maj. Erwin Orye, NATO CCDCOE
- Dr Anna-Maria Osula, Tallinn University of Technology, Estonia
- Dr Nikolas Ott, Microsoft, Belgium
- Capt. Barış Egemen Özkan, Turkish Naval Forces, Turkey
- Dr Piroska Páll-Orosz, Ministry of Defence, Hungary
- Piret Pernik, NATO CCDCOE
- Mauno Pihelgas, NATO CCDCOE
- Dr Narasimha Reddy, Texas A&M University, United States
- Lt. Col. Kurt Sanger, Department of Defense, United States
- Lt. Col. Massimiliano Signoretti, NATO CCDCOE
- Dr Max Smeets, ETH Zurich, Switzerland
- Prof. Edward Sobiesk, Army Cyber Institute, United States
- Dr Tim Stevens, King's College London, United Kingdom
- Maj. Damjan Štrucl, NATO CCDCOE
- Dr Jens Tölle, Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie, Germany
- Maria Tolppa, NATO CCDCOE
- Dr Risto Vaarandi, Tallinn University of Technology, Estonia
- Ann Väljataga, NATO CCDCOE
- Matthijs Veenendaal, Ministry of Defence, Netherlands

- Dr Adrian Venables, Tallinn University of Technology, Estonia
- Maj. Gábor Visky, NATO CCDCOE
- Prof. Sean Watts, School of Law, Creighton University, United States
- Dr Laurin Weissinger, Tufts University, United States
- Dr Christopher Whyte, Virginia Commonwealth University, United States
- Cmdr. Michael Widmann, NATO CCDCOE
- Jan Wünsche, NATO CCDCOE
- Philippe Zotz, NATO CCDCOE

Covid-19 and the Cyber Pandemic: A Plea for International Law and the Rule of Sovereignty in Cyberspace

François Delerue
Research Fellow
Institut de recherche stratégique de l'École militaire (IRSEM)
Paris, France
francois.delerue@alumni.eui.eu

Abstract: There has been an important increase in threats and attacks in cyberspace during the Covid-19 crisis. Incidentally, States and other actors have condemned this *cyber pandemic* and highlighted the incompatibility of these behaviours with international law and the framework of responsible State behaviour.

From the perspective of international law, the rule of sovereignty appears to have a central role to play in addressing the malicious cyber activities that have taken advantage of the coronavirus pandemic. Indeed, most of these malicious cyber activities may only constitute breaches of sovereignty. Sovereignty is, however, among the most unsettled and contentious parts of international law, even among the so-called 'like-minded' States, which have expressed very different interpretations.

Building on these observations, the present article investigates the different types of cyber operations that unfolded during the Covid-19 pandemic and questions their characterization in relation to the rules and principles of international law. It assesses the theoretical role of the rule of sovereignty in crisis management during a cyber pandemic as well as its actual use in State practice. Ultimately, it demonstrates the centrality of this rule of international law and how the current sanitary crisis may constitute a plea for its application – or perhaps its rejuvenation – and for its further development in State practice.

Keywords: *Covid-19, coronavirus, international law, sovereignty, espionage, SolarWinds*

1. INTRODUCTION

States and other actors have condemned the Covid-19 cyber pandemic and highlighted the incompatibility of such behaviours with international law and with the framework of responsible State behaviour. Cyber threats fuelled by Covid-19 were notably discussed during two Arria-Formula meetings of the United Nations Security Council. The first, which took place on 22 May 2020, focused on *Cyber Stability, Conflict Prevention and Capacity Building* and was organized by Estonia, in cooperation with Belgium, the Dominican Republic, Indonesia and Kenya.[1] The second meeting, which occurred on 26 August 2020, was dedicated to *Cyber Attacks Against Critical Infrastructure*, and was organized by Indonesia, in cooperation with Belgium, Estonia and Vietnam, as well as the International Committee of the Red Cross.[2] Representatives of different States spoke at these Arria-Formula meetings and reaffirmed the importance of international law in the fight against the cyber pandemic. The United States representatives at these two Arria-Formula meetings of the UN Security Council, for instance, condemned these behaviours and recalled the importance of international law.[3] Moreover, some States condemned these behaviours in their contributions to the ongoing UN processes on the peace and stability of cyberspace.

In addition to these collective efforts, States have also unilaterally condemned the cyber operations that took advantage of the Covid-19 pandemic and those that targeted institutions involved in the management of the crisis. In condemning them, they generally reasserted the centrality of international law in ensuring the peace and stability of cyberspace, including in these difficult times. For instance, the European Union condemned the malicious cyber activities exploiting the coronavirus pandemic through a declaration by the vice-president of the European Commission, Josep Borrell, on 30 April 2020. In it, he 'call[ed] upon every country to exercise due

---

[1] 'Arria-Formula Meeting: Cyber Stability, Conflict Prevention and Capacity Building' (*What's in blue*, 21 May 2020) <https://www.whatsinblue.org/2020/05/arria-formula-meeting-cyber-stability-conflict-prevention-and-capacity-building.php> accessed 24 March 2021.

[2] 'Arria-Formula Meeting on Cyber-Attacks Against Critical Infrastructure' (*What's in blue*, 25 August 2020) <https://www.whatsinblue.org/2020/08/arria-formula-meeting-on-cyber-attacks-against-critical-infrastructure.php> accessed 24 March 2021.

[3] United States Mission to the United Nations, Ambassador Cherith Norman Chalet, 'Remarks at a UN Security Council Arria-Formula Meeting on Cyber Stability and Responsible State Behavior in Cyberspace (via VTC)' (United States Mission to the United Nations 2020) <https://usun.usmission.gov/remarks-at-a-un-security-council-arria-formula-meeting-on-cyber-stability-and-responsible-state-behavior-in-cyberspace-via-vtc/> accessed 24 March 2021; United States Mission to the United Nations, Rodney Hunter, 'Remarks at a UN Security Council Arria-Formula Meeting on Cyber Attacks Against Critical Infrastructure (via VTC)' (United States Mission to the United Nations 2020) <https://usun.usmission.gov/remarks-at-a-un-security-council-arria-formula-meeting-on-cyber-attacks-against-critical-infrastructure-via-vtc/> accessed 24 March 2021.

diligence and take appropriate actions against actors conducting such activities from its territory, consistent with international law …'.[4]

Interestingly, however, we can observe a discrepancy between these general declarations and the condemnations in which the same States have denounced particular cyber operations that took advantage of the sanitary crisis. The United States, for instance, condemned the cyber operations that targeted a hospital in the Czech Republic in April 2020[5] and the Georgian Ministry of Health in September 2020.[6] Each time, they mentioned the 'framework of responsible State behavior in cyberspace, including nonbinding norms' but without making any reference to international law, nor stating which rule or principle of international law had been breached by these malicious activities. It is also conceivable that the United States considered these behaviours to be lawful and condemned them as unfriendly acts. These behaviours are likely to constitute violations of sovereignty, but their consequences were unlikely to have met the threshold of harm required by the United States as a criterion of a violation of sovereignty in cyberspace.[7]

Building on these observations, the present article explores the different types of cyber operations associated with the Covid-19 pandemic and questions their characterization in relation to existing rules and principles of international law. It assesses the theoretical role of the rule of sovereignty in the management of the cyber pandemic crisis, as well as its actual application and implementation in State practice. Ultimately, it demonstrates the centrality of this rule of international law and how the current sanitary crisis may constitute a plea for its application, if not for its rejuvenation, but also for its further development in State practice.

There are five sections in this article, the introduction being the first. The second section analyses the different types of cyber operations associated with the Covid-19 pandemic. The third section briefly introduces the international law applicable to cyber operations. The fourth section assesses the lawfulness of the cyber pandemic under international law. Finally, the fifth section discusses the role of the rule of sovereignty

---

4   European Union, 'Declaration by the High Representative Josep Borrell, on Behalf of the European Union, on Malicious Cyber Activities Exploiting the Coronavirus Pandemic' (Council of the European Union, 30 April 2020) <https://www.consilium.europa.eu/en/press/press-releases/2020/04/30/declaration-by-the-high-representative-josep-borrell-on-behalf-of-the-european-union-on-malicious-cyber-activities-exploiting-the-coronavirus-pandemic/> accessed 24 March 2021.

5   United States Secretary of State, Michael R. Pompeo, 'The United States Concerned by Threat of Cyber Attack Against the Czech Republic's Healthcare Sector' (U.S. Department of State, 17 April 2020) <https://cz.usembassy.gov/the-united-states-concerned-by-threat-of-cyber-attack-against-the-czech-republics-healthcare-sector/> accessed 24 March 2021.

6   United States Embassy in Georgia, 'U.S. Embassy Statement on September 1, 2020 Cyberattack against Georgian Ministry of Health' (U.S. Embassy in Georgia, 1 September 2020) <https://ge.usembassy.gov/u-s-embassy-statement-on-september-1-2020-cyberattack-against-georgian-ministry-of-health/> accessed 24 March 2021.

7   United States, Brian J. Egan, 'Remarks on International Law and Stability in Cyberspace' (US Department of State 2016) <https://2009-2017.state.gov/s/l/releases/remarks/264303.htm> accessed 24 March 2021.

in managing the cyber pandemic crisis and how it may affect the different approaches adopted by some States in interpreting this rule of international law.

2. DECONSTRUCTING THE CYBER PANDEMIC

The Covid-19 pandemic has been marked by an important increase in the number of threats and operations in cyberspace. This cyber pandemic takes mainly two forms: first, some cyber threats have taken advantage of the pandemic-induced crisis; second, other cyber threats have been expressly directed at the health care sector and at the institutions involved in the management of the crisis. Among others, some States are believed to be responsible for a certain portion of these malicious cyber activities. The objective of the present section is to briefly introduce these different cyber operations and to identify which of them may have been conducted or sponsored by States and, incidentally, the rules and principles of international law that may be applicable in such cases.[8] Aside from the cyber pandemic, Covid-19 has also been accompanied by an *infodemic*; that is to say, disinformation campaigns that use the pandemic as a vector. Because the present article focuses on cyber operations, the infodemic lies outside its scope and is not studied here.[9]

The first category covers cyber threats that take advantage of the pandemic and may be qualified as opportunistic cyber operations. The spread of Covid-19 has been marked by an exponential digitalization of our lives, either for work, education or entertainment, or in our interactions with loved ones. Moving these activities online has created numerous new vulnerabilities that may be exploited by malicious actors. The fact that many workers have been working remotely – thus, shifting their activities to personal computers and networks that may not have the same security features as the ones usually used at the office – is also a source of vulnerability. The European Union Agency for Cybersecurity (ENISA),[10] Europol,[11] Interpol,[12] and

---

[8] On the question of the attribution of cyber operations, see generally: François Delerue, *Cyber Operations and International Law* (Cambridge University Press 2020) 55–189; Dennis Broeders, Els De Busser and Patryk Pawlak, 'Three Tales of Attribution in Cyberspace. Criminal Law, International Law and Policy Debates' (The Hague Program for Cyber Norms, Policy Brief 2020) <https://www.thehaguecybernorms.nl/research-and-publication-posts/three-tales-of-attribution-in-cyberspace-criminal-law-international-law-and-policy-debates> accessed 24 March 2021; Kristen E Eichensehr, 'The Law and Politics of Cyberattack Attribution' (2020) 67 U.C.L.A. Law Review 520, 520–598; Michael N Schmitt and Liis Vihul (eds), *The Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (2nd edn, Cambridge University Press 2017) 87–100.

[9] See notably: Barrie Sander and Nicholas Tsagourias, 'The Covid-19 Infodemic and Online Platforms as Intermediary Fiduciaries under International Law' (2020) 11 JHLS 331, 331–347; Marko Milanovic and Michael N Schmitt, 'Cyber Attacks and Cyber (Mis)Information Operations During a Pandemic' (2020) 11 JNSLP 247, 266 et seq.

[10] ENISA, 'COVID-19' (European Union Agency for Cybersecurity 2021) <https://www.enisa.europa.eu/topics/wfh-covid19> accessed 24 March 2021.

[11] Europol, 'COVID-19 Sparks Upward Trend in Cybercrime' (Europol, 5 October 2020) <https://www.europol.europa.eu/newsroom/news/covid-19-sparks-upward-trend-in-cybercrime> accessed 24 March 2021.

[12] INTERPOL, 'COVID-19 Cyberthreats' (INTERPOL 2021) <https://www.interpol.int/Crimes/Cybercrime/COVID-19-cyberthreats> accessed 24 March 2021.

the United States Cybersecurity and Infrastructure Agency (CISA)[13] – among many others – have drawn attention to these cyber threats. Notably, they pointed out that cybercriminals have been using the pandemic as a vector for phishing campaigns, ransomware attacks, and for spreading malware, online scams and disinformation campaigns. As cybercriminal activities are outside the scope of this article, they will not be further discussed.

In addition, the digitalization of the life of citizens throughout the world may have been exploited by some States. State agents and actors operating as their proxies may be using similar techniques, notably phishing campaigns, to take advantage of the vulnerabilities that arose from the digitalization of our societies. In weakening the cyber hygiene of individuals, especially as they continue working from home on personal devices and using less secure networks, the pandemic increases the potential for attacks and creates new opportunities for malicious actors to target these individuals. In doing so, the main objective is likely to gain access to the credentials of the targeted individuals and, ultimately, access to their devices to steal, compromise or destroy data.

Furthermore, the second category deals with cyber operations that target actors involved in the management of the Covid-19 crisis. The healthcare sector, in particular, faces numerous threats from cyberspace while they need to treat patients suffering from the coronavirus.[14] For instance, hospitals in various countries have been targeted by different cyber threats, such as ransomware attacks and Distributed Denial of Service (DDoS) attacks.[15]

Hence, in these challenging times, information is key. It appears that different actors have conducted cyber operations to get access to information and data on the spread of the virus and on the measures adopted in different countries. The Chinese cybersecurity company Qihoo 360 accused the advanced persistent threat (APT) known as DarkHotel, allegedly linked to South Korea, of having conducted a cyber espionage campaign against Chinese and international institutions, presumably to obtain information on the spread of the virus.[16] Similarly, APT 32, also known as OceanLotus Group, a group generally believed to be linked to Vietnam, has been

---

[13]   CISA, 'Coronavirus' (United States Cybersecurity and Infrastructure Agency 2021) <https://www.cisa.gov/coronavirus> accessed 24 March 2021.

[14]   Liviu Arsene, '5 Times More Coronavirus-Themed Malware Reports during March' (*Bitdefender*, 20 March 2020) <https://labs.bitdefender.com/2020/03/5-times-more-coronavirus-themed-malware-reports-during-march/> accessed 24 March 2021.

[15]   Matt Burgess, 'Hackers Are Targeting Hospitals Crippled by Coronavirus' (*Wired*, 22 March 2020) <https://www.wired.co.uk/article/coronavirus-hackers-cybercrime-phishing> accessed 24 March 2021; Emmanuel Paquette, 'En pleine crise du coronavirus, les hôpitaux de Paris victimes d'une cyberattaque' (*L'Express*, 23 March 2020) <https://lexpansion.lexpress.fr/high-tech/en-pleine-crise-du-coronavirus-les-hopitaux-de-paris-victimes-d-une-cyberattaque_2121692.html> accessed 24 March 2021.

[16]   Jeff Stone, 'A Chinese Security Firm Says DarkHotel Hackers Are behind an Espionage Campaign, but Researchers Want More Details' (*CyberScoop*, 6 April 2020) <https://www.cyberscoop.com/dark-hotel-qihoo-360-covid-19/> accessed 24 March 2021.

accused of having conducted cyber espionage activities against the staff of the Chinese Ministry of Emergency Management and of the Government of Wuhan.[17] At the global level, international organizations involved in the management of the sanitary crisis and the exchange of information have also been targeted.[18] The staff of the World Health Organization, for instance, has been targeted by phishing email campaigns.[19]

Additionally, the race for a vaccine against Covid-19 has been subjected to cyber operations targeting research institutions. Canada, the United Kingdom, and the United States have accused APT 29, also known as Cozy Bear, a group generally believed to be associated with Russian intelligence agencies, of using malware named WellMess or WellMail to target institutions involved in the development of Covid-19 vaccines.[20] Likewise, APT 38, also known as the Lazarus Group, and believed to be linked to North Korea, has been accused of targeting a pharmaceutical company developing a Covid-19 vaccine as well as a government institution involved in the management of the crisis.[21]

To sum up, the cyber operations in the second group show two different trends. On the one hand, some cyber operations aim at disrupting the daily management of hospitals; these activities normally do not match the usual profile of State conducted or sponsored operations. On the other hand, certain cyber operations strive to gather information on the spread of the virus, the management of the crisis by different actors, as well as to gain access to research on the development of a vaccine; the latter are more likely to be conducted or sponsored by States.

In conclusion, this section assessed the malicious cyber activities linked to the Covid-19 pandemic and showed that States are likely to conduct or sponsor operations to gather information and data, either targeting individuals that are more vulnerable in these challenging times or institutions involved in the management of the crisis and in the development of vaccines. The identification of the types of cyber operations that may have been conducted by States and their proxies allows us to assess their lawfulness

17   Raphael Satter and Jack Stubbs, 'Vietnam-Linked Hackers Targeted Chinese Government over Coronavirus Response: Researchers' (Reuters, 22 April 2020) <https://www.reuters.com/article/us-health-coronavirus-cyber-vietnam/vietnam-linked-hackers-targeted-chinese-government-over-coronavirus-response-researchers-idUSKCN2241C8> accessed 24 March 2021.

18   Kaspersky Lab (GReAT), 'APT Annual Review: What the World's Threat Actors Got up to in 2020' (*Securelist*, 3 December 2020) <https://securelist.com/apt-annual-review-what-the-worlds-threat-actors-got-up-to-in-2020/99574/> accessed 24 March 2021.

19   Joseph Menn and others, 'Hackers Linked to Iran Target WHO Staff Emails during Coronavirus' (Reuters, 2 April 2020) <https://www.reuters.com/article/us-health-coronavirus-cyber-iran-exclusi-idUSKBN21K1RC> accessed 24 March 2021.

20   UK NCSC, 'Advisory: APT29 Targets COVID-19 Vaccine Development' (United Kingdom's National Cyber Security Centre (NCSC) 2020) <https://www.ncsc.gov.uk/files/Advisory-APT29-targets-COVID-19-vaccine-development.pdf> accessed 24 March 2021.

21   Seongsu Park, 'Lazarus Covets COVID-19-Related Intelligence' (*Securelist*, 23 December 2020) <https://securelist.com/lazarus-covets-covid-19-related-intelligence/99906/> accessed 24 March 2021.

(Section 4). But, before that, the next section briefly introduces the international legal framework applicable to cyber operations.

3. INTERNATIONAL LAW APPLIES TO CYBER OPERATIONS

International law, and in particular the Charter of the United Nations, is the backbone of contemporary international relations and remains crucial in maintaining international peace and security. Nowadays, the applicability of international law to cyberspace is consensual among States and other actors: international law applies to cyberspace and cyber operations.[22] This has notably been affirmed by the consensual reports of the United Nations Group of Governmental Experts (GGE) on Developments in the Field of Information and Telecommunications in the Context of International Security in 2013 and 2015, and later confirmed by the majority of States on various occasions.[23] The question of the applicability of international law being settled, the debate has moved on to the question of how the rules and principles of international law are to be applied to cyberspace.

At the multilateral level, the effort to clarify the interpretation of the rules and principles of international law has already been undertaken by the third, fourth and fifth UN GGEs. The failure of the fifth UN GGE, in June 2017, actually resulted from this endeavour as it highlighted certain divergences among the participating States. The disagreement that erupted between the participating experts of the fifth UN GGE had nothing to do with the applicability of certain branches of international law to cyberspace but rather with the opportunity to enshrine a specific interpretation in the

---

[22] See notably: Heather Harrison Dinniss, *Cyber Warfare and the Laws of War* (Cambridge Studies in International and Comparative Law, Cambridge University Press 2012); Georg Kerschischnig, *Cyberthreats and International Law* (Eleven International Publishing 2012); Michael N Schmitt (ed), *The Tallinn Manual on the International Law Applicable to Cyber Warfare* (Cambridge University Press 2013); Katharina Ziolkowski (ed), *Peacetime Regime for State Activities in Cyberspace: International Law, International Relations and Diplomacy* (NATO Cooperative Cyber Defence Centre of Excellence 2013); Marco Roscini, *Cyber Operations and the Use of Force in International Law* (Oxford University Press 2014); Scott J Shackelford, *Managing Cyber Attacks in International Law, Business, and Relations: In Search of Cyber Peace* (Cambridge University Press 2014); Johann-Christoph Woltag, *Cyber Warfare: Military Cross-Border Computer Network Operations under International Law* (Intersentia 2014); Yaroslav Radziwill, *Cyber-Attacks and the Exploitable Imperfection of International Law* (Brill & Martinus Nijhoff Publishers 2015); Schmitt and Vihul (n 8); Henning Lahmann, *Unilateral Remedies to Cyber Operations: Self-Defence, Countermeasures, Necessity, and the Question of Attribution* (Cambridge University Press 2020); Delerue (n 8).

[23] See, for instance: UNGA 'Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security' (24 June 2013) UN Doc A/68/98 2013 8, para 19; UNGA 'Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security' (22 July 2015) UN Doc A/70/174 2015 12, para 24 et seq.

final report, as well as the particularities of interpreting them in the cyber context.[24] Today, these questions are again part of the mandate of the ongoing sixth UN GGE and of the Open-Ended Working Group (OEWG) on developments in the field of information and telecommunications in the context of international security, created by UNGA Resolutions 73/266 and 73/27, respectively.

In recent years, there has also been important evolution in State practices regarding the international law applicable to cyberspace, in two main directions.

First, a growing number of States have publicized their approach on the rules and principles of international law applicable to cyberspace.[25] Two important caveats must be addressed though. On the one hand, fewer than a dozen States have made their interpretation public. On the other hand, the vast majority of the detailed approaches now publicized have been released by Western States. Therefore, the picture we have is geographically limited and partial. This second limitation may, however, recede in the future for two reasons. First, the recent publication by Iran of its approach may actually incentivize other non-Western States to follow suit.[26] It is indeed the first time that a non-Western State has publicly disclosed a detailed approach on this matter. Second, UNGA Resolution 73/266 requested that the States participating in the UN GGE submit their views on how international law should be applied to cyberspace. As such, there is a growing push for the States not taking part in the UN GGE to disclose their views as well, notably within the framework of the OEWG.

Second, States are increasingly developing and strengthening their practice on conducting and reacting to cyber operations. A growing number of States has been integrating cyber-related dispositions in their military manuals and domestic regulations on military and intelligence activities, a process that reflects, to some extent, their compliance with their international legal obligations. Yet, it is difficult to assess the compliance of their practice in conducting or sponsoring cyber operations, since it remains a predominantly covert practice. As for reacting to cyber operations, some States have developed a practice of 'naming and shaming' those States responsible for conducting malicious cyber activities. An important limitation to this observation is that these public attributions have only been done by a limited number of States,

---

24   François Delerue, Frédérick Douzet and Aude Géry, *The Geopolitical Representations of International Law in the International Negotiations on the Security and Stability of Cyberspace / Les Représentations Géopolitiques Du Droit International Dans Les Négociations Internationales Sur La Sécurité La Stabilité Du Cyberespace* (IRSEM and EU Cyber Direct 2020) <https://eucyberdirect.eu/content_research/the-geopolitical-representations-of-international-law-in-the-international-negotiations-on-the-security-and-stability-of-cyberspace/> accessed 24 March 2021.

25   See the analysis in Przemysław Roguski, 'Application of International Law to Cyber Operations: A Comparative Analysis of States' Views' (Policy Brief, The Hague Program for Cyber Norms 2020) <https://www.thehaguecybernorms.nl/news-and-events-posts/policy-brief-application-of-international-law-to-cyber-operations-a-comparative-analysis-of-states-views> accessed 24 March 2021.

26   Iran, 'General Staff of Iranian Armed Forces Warns of Tough Reaction to Any Cyber Threat' (NOURNEWS Analytics & News Agency 2020) <https://nournews.ir/En/News/53144/General-Staff-of-Iranian-Armed-Forces-Warns-of-Tough-Reaction-to-Any-Cyber-Threat> accessed 24 March 2021.

usually Western ones, and predominantly by the Five Eyes Member States (Australia, Canada, New Zealand, the United Kingdom and the United States).[27] Interestingly, the vast majority of cases of the public attribution or condemnation of cyber operations have made no reference to international law. Only a few have come out to make loose references to international law or to the international rules-based order. None of these statements has ever clearly characterized which rule or principle of international law has been breached, nor referred to the categories of the international legal framework used to attribute and react to these acts.

Yet, other actors have been active in clarifying how international law applies to cyberspace and cyber operations. The most advanced example is the *Tallinn Manual* process initiated in 2009 by the NATO Cooperative Cyber Defence Centre of Excellence (NATO CCDCOE), which led to the publication of the *Tallinn Manual on the International Law Applicable to Cyber Warfare (Tallinn Manual 1.0)* in 2013[28] and the *Tallinn Manual on the International Law Applicable to Cyber Operations (Tallinn Manual 2.0)* in 2017.[29] The NATO CCDCOE just announced the beginning of the work on a third version of the *Tallinn Manual*.[30] Another good example is the Cyber Law Toolkit, which offers an in-depth exemplification of the application of international law to cyber operations through scenarios.[31] Other actors, including some from the private sector, NGOs and expert groups, have addressed the questions pertaining to international law as part of the broader theme of the framework of responsible State behaviour – a framework that also includes norms of responsible behaviour and confidence-building measures. Moreover, recent initiatives and developments have demonstrated that international law applies, and offers a relevant legal framework, to the cyber operations that take advantage of the sanitary crisis, such as the Oxford Process.[32] Additionally, different academic publications have come to the same conclusion, such as, for instance, the seminal article by Marko Milanovic and Michael N. Schmitt.[33]

---

[27] Florian J Egloff, 'Contested Public Attributions of Cyber Incidents and the Role of Academia' (2020) 41 Contemporary Security Policy 55, 61.

[28] Schmitt (n 22).

[29] Schmitt and Vihul (n 8).

[30] 'CCDCOE to Host the Tallinn Manual 3.0 Process' (NATO Cooperative Cyber Defence Centre of Excellence, 14 December 2020) <https://ccdcoe.org/news/2020/ccdcoe-to-host-the-tallinn-manual-3-0-process/> accessed 24 March 2021.

[31] 'International Cyber Law in Practice: Interactive Tookit' (Cyber Law Toolkit) <https://cyberlaw.ccdcoe.org/> accessed 24 March 2021.

[32] Two online events gathered international lawyers to debate the rules and principles of international law applicable in such circumstances and led to the adoption of related statements: 'The Oxford Statement on the International Law Protections Against Cyber Operations Targeting the Health Care Sector' (Oxford Institute for Ethics, Law and Armed Conflict (ELAC), University of Oxford 2020) <https://elac.web.ox.ac.uk/the-oxford-statement-on-the-international-law-protections-against-cyber-operations-targeting-the-hea> accessed 24 March 2021; 'The Second Oxford Statement on International Law Protections of the Healthcare Sector During Covid-19: Safeguarding Vaccine Research' (Oxford Institute for Ethics, Law and Armed Conflict (ELAC), University of Oxford 2020) <https://elac.web.ox.ac.uk/article/the-second-oxford-statement> accessed 24 March 2021.

[33] Milanovic and Schmitt (n 9).

This brief introduction to the debates surrounding the international law applicable to cyber operations leads to three observations. First, there is no contestation of the international legal framework applicable to cyber operations: the rules and principles of international law do apply to cyber operations. As highlighted regarding the failure of the fifth UN GGE in 2017, the disagreement is mainly political rather than legal. It did not show any opposition to the applicability of the rules and principles nor to their interpretation. Second, the international discussions, the unilateral statements by States on their respective approaches, the scholarly literature and all the other initiatives provide us with a good picture of the relevant rules and principles that are applicable to cyber operations, including in these challenging times of the current pandemic. However, the implementation of the international legal framework in State practice remains relatively limited. Third, despite the absence of opposition to the international legal framework, some divergences appear on its interpretation and on the concrete application of certain rules and principles. In fact, the interpretation of the rule or principle of sovereignty appears to be the most contentious issue, as shown in the next section.

4. APPLYING INTERNATIONAL LAW TO THE CYBER PANDEMIC

In this section, the objective is to assess whether the cyber operations conducted or sponsored by States during the pandemic constitute internationally wrongful acts. To be an internationally wrongful act, the action or omission must be attributable to a State and constitute a breach of an international obligation.[34] The question of attribution is not discussed in the present article and we will focus on the second element.[35] There are three main obligations that may be breached by cyber operations in general: the prohibition of the use or threat of force, the prohibition of intervention, and the rule of sovereignty.[36] In addition, the principle of due diligence appears to be particularly relevant in addressing cyber threats related to the Covid-19 pandemic.[37]

In a recent article, Marko Milanovic and Michael N. Schmitt assessed that the majority of cyber operations against healthcare facilities and capabilities may be violating the sovereignty of other States.[38] I agree with this assessment and this will be demonstrated in the present section. As discussed earlier, in taking advantage of the Covid-19

---

[34] *Articles on Responsibility of States for Internationally Wrongful Acts* (adopted by the International Law Commission at its fifty-third session in 2001, annexed to General Assembly Resolution 56/83 of 12 December 2001, and corrected by Document A/56/49 (Vol I)/Corr4), Article 2. For a discussion on the characterization of cyber operations as internationally wrongful acts, see: Schmitt and Vihul (n 8) 84, rule 14; Delerue (n 8) 381.

[35] Delerue (n 8) 55–189.

[36] Schmitt and Vihul (n 8), rules 4, 66, 68–70; Delerue (n 8) 193–342.

[37] François Delerue and Joanna Kulesza, 'Cybersecurity in the Year of the Plague: Due Diligence as a Remedy to Malicious Activities' (2020) 2 Tecnologie e Diritto 404, 404–419.

[38] Milanovic and Schmitt (n 9).

pandemic, several States and their proxies have been predominantly conducting cyber operations aimed at gathering information and data, either by targeting individuals that are more vulnerable in these challenging times or institutions involved in the management of the crisis and in the race to find a vaccine. This section analyses these cyber operations in relation to the main rules and principles of the international law applicable to cyber operations.

A. The Cyber Pandemic and the Prohibition of the Use of Force

To constitute an unlawful use of force, a cyber operation would need to provoke physical damage, human injury or death.[39] There is no agreement on whether a cyber operation with no physical effect, but causing very significant damage in cyberspace, may amount to unlawful use of force.[40]

It is conceivable that some cyber operations taking advantage of the Covid-19 pandemic could have significant consequences and thus be characterized as unlawful uses of force. For instance, we could consider the example of a State-sponsored ransomware disrupting the normal running of a hospital, thus leading to the death of patients who could not receive the necessary care in time or because they received the wrong treatment.[41] That being said, none of the alleged State conducted or sponsored cyber operations that have occurred since the outbreak of Covid-19 came close to this required threshold of consequences. Therefore, even if it is theoretically possible, it seems highly unlikely that State conducted or sponsored cyber operations taking advantage of Covid-19 would constitute a use of force.

B. The Cyber Pandemic and the Prohibition of Intervention

To constitute an unlawful intervention, a cyber operation must meet three criteria, as stressed most famously by the International Court of Justice in the *Nicaragua* case.[42] First, an intervention must be carried out by a State or its proxy acting against another State. Second, the prohibited intervention concerns matters in which the targeted State is permitted to decide freely, encompassing external or internal affairs. Third, the element of coercion constitutes an essential component of a prohibited intervention.

---

39   Schmitt and Vihul (n 8) 329–338.
40   For instance, the French ministry of defence stated that 'France does not rule out the possibility that a cyberoperation without physical effects may also be characterized as a use of force', in: France, 'International Law Applied to Operations in Cyberspace' (ministère des Armées 2019) 7 <https://www.defense.gouv.fr/content/download/567648/9770527/file/international+law+applied+to+operations+in+cyberspace.pdf> accessed 24 March 2021.
41   In September 2020, a ransomware attack, not attributed to a State, that targeted a hospital in Düsseldorf was believed to have contributed to the death of a patient by delaying her treatment. The subsequent investigation concluded, however, that the ransomware was not responsible for the death. Willian Ralston, 'The untold story of a cyberattack, a hospital and a dying woman' (*Wired UK*, 11 November 2020) <https://www.wired.co.uk/article/ransomware-hospital-death-germany> accessed 24 March 2021.
42   *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v United States of America)* (Merits) [1986] ICJ Rep 14, 107–108, para 205.

A prohibited intervention must constitute an attempt to coerce the targeted State by directly or indirectly interfering in the internal or external affairs of this State.[43]

The first two criteria are not specifically challenged by the features of the above-discussed cyber operations and are not discussed further for that reason. Conversely, assessing whether these cyber operations meet the third criterion is a trickier question. Indeed, the vast majority of cyber operations observed during the Covid-19 pandemic aimed at collecting data and information but did not have a coercive objective. The objective being to gather data and information to support the sponsoring State's policy and strategy and not to influence the targeted State.

True, the stolen data may be leaked or instrumentalized to coerce the targeted State. Yet, in such cases, the theft and the use of the data are two different acts,[44] the former being likely to constitute a breach of sovereignty while the second being more likely to be an unlawful intervention.

C. The Cyber Pandemic and Sovereignty

Cyber malicious acts taking advantage of the Covid-19 pandemic may, in most cases, constitute a violation of the sovereignty of the targeted States.[45] Indeed, most of these cyber operations aimed at penetrating computer systems and networks located on the territory of other States are meant to access and steal data. Unauthorized penetration into computer systems constitutes the basis of a violation of sovereignty. Yet, it must be noted that it remains one of the most contentious questions dealing with the international law applicable to cyber operations, since the States have adopted very different approaches, which I summarize below.

The different approaches revolve around three mains debates about sovereignty in cyberspace. First, whether sovereignty is a rule or a principle of international law. Second, on the reach of sovereignty when it is applied to cyberspace. Third, there remains a plurality of views on what may constitute a breach of territorial sovereignty in cyberspace.

First, the nature of territorial sovereignty in cyberspace is not settled. Sovereignty is a general principle of international law from which certain rules are derived, including the prohibition of the violation of territorial sovereignty.[46] Both rules and principles are sources of international law, and they are notably listed in Article 38 of the Statute

---

43 Philip Kunig, 'Intervention, Prohibition Of', *MPEPIL* (2008), para 1; Gaetano Arangio-Ruiz, 'Human Rights and Non-Intervention in the Helsinki Final Act' (1977) 157 RCADI 195, 257, 261 *et seq*.
44 Delerue (n 8) 241–256.
45 Milanovic and Schmitt (n 9) 252–256.
46 In outlining the Israeli perspective on the international law applicable to cyber operations, Roy Schöndorf wrote an interesting analysis of these different aspects of sovereignty in cyberspace: 'Israel's Perspective on Key Legal and Practical Issues Concerning the Application of International Law to Cyber Operations' (9 December 2020) <https://www.ejiltalk.org/israels-perspective-on-key-legal-and-practical-issues-concerning-the-application-of-international-law-to-cyber-operations/> accessed 24 March 2021.

of the International Court of Justice.[47] Rules refer to the actual norms of international law, from treaties or customary international law, for example. Furthermore, principles refer to the more abstract notions from which rules flow. While States agree on the existence of a general principle of sovereignty, they have divergent opinions on the rules flowing from that principle. Indeed, while some consider sovereignty only as a principle of international law in the cyber realm (e.g. the United Kingdom),[48] the majority argues that it is a rule.

Second, there is no consensus on what constitutes State sovereignty in cyberspace. For instance, there are ongoing debates over whether States are entitled to exercise sovereignty over data located on computers belonging to other entities which may or may not be located on the State's territory.[49] The confusion is amplified by the conflation between sovereignty as a political concept and sovereignty as defined by international law.

Third, there are multiple definitions of what may amount to a breach of territorial sovereignty when it comes to cyber operations. Among the limited number of States that have publicly disclosed their views on the matter, we can identify three main perspectives. In the first approach, any cyber operation that penetrates a foreign system or produces effects over it constitutes a violation of sovereignty. This is, for instance, the French approach.[50] Then, in the second approach, a cyber operation penetrating a foreign system constitutes a violation of sovereignty only if it meets a certain threshold of harm. This is the approach adopted in the *Tallinn Manual 2.0*[51] and by the United States.[52] It should be noted, however, that the position expressed recently by Paul Ney,[53] the General Counsel of the US Department of Defence, seemed to lean towards a third approach.[54] With that last approach, territorial sovereignty cannot be breached by a cyber operation unless it constitutes a violation of the principle of non-intervention. This is, for instance, the British approach.[55] These three different approaches have been formulated by Western States, usually considered to be 'like-minded' States, and it is plausible that other approaches may be expressed by other States in the future.

---

[47]   *Statute of the International Court of Justice*, annexed to the Charter of the United Nations, adopted 26 June 1945, entered into force 24 October 1945, 3 Bevans 1179, 59 Stat. 1031, T.S. 993, 39 AJIL Supp. 215 (1945).

[48]   United Kingdom, Jeremy Wright, 'Cyber and International Law in the 21st Century' (UK Attorney General's Office 2018) <https://www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century> accessed 24 March 2021.

[49]   See, for instance, the discussion in: Roy Schöndorf (n 46).

[50]   France (n 40).

[51]   Schmitt and Vihul (n 8) 17–26, rule 4.

[52]   United States, Brian J. Egan (n 7).

[53]   Paul C Ney, Jr, 'DOD General Counsel Remarks at U.S. Cyber Command Legal Conference' (2020) <https://www.defense.gov/Newsroom/Speeches/Speech/Article/2099378/dod-general-counsel-remarks-at-us-cyber-command-legal-conference/> accessed 24 March 2021.

[54]   Michael N Schmitt, 'The Defense Department's Measured Take on International Law in Cyberspace' (*Just Security*, 11 March 2020) <https://www.justsecurity.org/69119/the-defense-departments-measured-take-on-international-law-in-cyberspace/> accessed 24 March 2021.

[55]   United Kingdom, Jeremy Wright (n 48).

If we apply the three approaches to the malicious cyber operations taking advantage of the Covid-19 pandemic, they would constitute violations of sovereignty under the first approach but be deemed lawful under the third approach. Moreover, it seems doubtful that these cyber operations met the threshold of harm required by the States having adopted the second approach.

Aside from these three approaches, the *Tallinn Manual 2.0*, and some States such as the Netherlands,[56] have laid out another basis that may constitute a breach of the rule of sovereignty: when 'there has been an interference or usurpation of inherently governmental functions'.[57] There are two criteria to sustain this one: first, it must concern 'inherently governmental functions'. As rightly pointed out by Marko Milanovic and Michael N. Schmitt, while the management of the sanitary crisis is likely to be considered an inherently governmental function, it is more debatable regarding the provision of healthcare.[58] Consequently, this first criterion needs to be assessed on a case-by-case basis. According to the second criterion, the concerned cyber operation should be an interference or usurpation of these functions. As previously highlighted, most cyber operations within that purview aim at accessing and stealing data, without further action. Even if this data is linked to inherently governmental functions, it appears debatable – if not unlikely – that they may be seen as either a usurpation or an interference of these functions.

In conclusion, most cyber operations taking advantage of the Covid-19 pandemic are likely to constitute, in theory, violations of the territorial sovereignty of the affected States, yet unlikely to be considered as such by several States under their own interpretation of the rule in this particular context. For the majority of States that have expressed their views on the international law applicable to cyber operations, these cyber operations would fall short of a violation of sovereignty, either because they did not cause sufficient harm, they did not interfere or usurp inherently governmental functions, or because they did not constitute unlawful interventions.

5. THE NECESSITY OF AN EVOLUTION OF THE STATES' APPROACH ON THE INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS

This assessment of the lawfulness of the cyber operations that take advantage of the Covid-19 pandemic confirms that the main challenge is not the identification of the relevant rules or principles of international law but rather their interpretation and

---

56 The Netherlands, 'Letter to the Parliament on the International Legal Order in Cyberspace (Appendix on International Law in Cyberspace)' (Government of the Netherlands 2019) 3 <https://www.government.nl/documents/parliamentary-documents/2019/09/26/letter-to-the-parliament-on-the-international-legal-order-in-cyberspace> accessed 24 March 2021.

57 Schmitt and Vihul (n 8) 20–23, paras 10, 15–18. See also the analysis of this basis in the context of cyber espionage, in: Russell Buchan, *Cyber Espionage and International Law* (Bloomsbury Publishing 2018) 61.

58 Milanovic and Schmitt (n 9) 253, 255–256.

implementation by States. This assessment also highlights that the cyber operations taking place in these challenging times are similar to the ones usually conducted or sponsored by States: they are predominantly activities of cyber espionage. The situation is different not because the cyber operations are different but because their number may have increased and, more importantly, because their targets have received increased attention. Therefore, interest in this topic is not linked to an evolution in State practice in the specific context of the Covid-19 pandemic but rather to an evolution in the way we apprehend these matters in these challenging times.

International law offers a legal framework that applies to and regulates such behaviours; it also provides response mechanisms for the injured States, such as countermeasures.[59] Yet, several States and scholars have decided to take an unconventional approach to the rule of sovereignty in cyberspace by either denying its existence or conditioning it to a threshold of harm. Why such a specific approach in the cyber realm? In any other domain, the mere unauthorized trespassing of a border, for instance by an aircraft or boat, is enough to constitute a violation of sovereignty and no threshold of harm is required. In cyberspace, the trespassing of a border is constituted by the unauthorized penetration into a computer system regardless of the potential harm caused.[60] It has been argued that the addition of a threshold of harm as well as the opposition to the existence of a rule of territorial sovereignty in cyberspace was motivated by the willingness of States to avoid limitations on their espionage capabilities. Adopting an approach that is too broad on the rule of sovereignty in cyberspace would indeed contradict espionage activities that heavily rely on the penetration of foreign computer systems.[61]

Building on these observations, it may be asserted that, by highlighting State practice in cyberspace, the Covid-19 cyber pandemic calls upon us to reconsider two questions, starting with the different approaches to the rule of sovereignty in cyberspace that coexist. Then, we need to reassess the difficult equilibrium between the necessity to ensure the peace and stability of cyberspace through international law and the framework of responsible State behaviour, and the willingness of States to pursue certain unfriendly, if not adversarial, activities, such as intelligence gathering campaigns.

In fact, it may be time for States to rethink their approach to the rule of sovereignty in cyberspace and to decide whether such activities (i.e. cyber espionage campaigns) should be deemed unlawful or not, according to their approaches to how international law applies in cyberspace. If they are to be considered lawful, States may continue to condemn them: despite their lawfulness, they could be deemed unethical or immoral. In that case, however, States would deprive themselves of the lawful responses offered

---

59   Schmitt and Vihul (n 8) 111–134, rules 20–25; Delerue (n 8) 433–460.
60   Delerue (n 8) 215–219.
61   On the international law applicable to cyber espionage, see generally: Asaf Lubin, 'The Liberty to Spy' (2020) 61 Harvard International Law Journal 185; Buchan (n 57).

by the law of countermeasures, which are useful and relevant tools to compel the wrongful State to cease its behaviour and repair eventual injuries.

The recent SolarWinds case and the US 'defend forward' cyber strategy lead to a similar questioning. First, in the SolarWinds case, countless articles and comments have argued that SolarWinds constitutes an armed attack and that the United States would be entitled to invoke their right of self-defence in response.[62] Yet, as rightly pointed out by Jack Goldsmith, this seems to be purely a cyber espionage campaign in which State-backed hackers penetrated computer systems to access and steal data.[63] In that sense, the SolarWinds case is very similar to several cases of cyber operations that took advantage of the Covid-19 pandemic. They are cyber espionage activities pure and simple. By restraining the rule of sovereignty in cyberspace, States have made such activities lawful and have thus deprived themselves of the responses allowed by international law. Second, the implementation of the 'defend forward' cyber strategy by the United States is likely to take the form of cyber operations breaching given rules and principles of international law, predominantly the rule of sovereignty. In 2019, for instance, the *New York Times* reported that the US Cyber Command hacked the computer systems running the Russian power grid as a preparatory measure for potential further actions.[64] Such behaviours, which are to some extent comparable to the actions against SolarWinds, are likely to constitute blatant violations of the rule of sovereignty. These examples highlight the discrepancy that may exist between rhetoric and practice for some States.

In disregarding certain rules of international law in practice, as well as in limiting their reach through a particular interpretation of international law, States appear to be turning their backs on the international rules-based order. Such an approach bears the risk of endangering the international peace and stability of cyberspace. If international law is not perfect and has not prevented breaches of peace and aggressions in the past, it constitutes a powerful tool and the best regulatory framework at our disposal if we want to avoid turning cyberspace into a new Wild West.

---

[62]   See for instance, Thomas P Bossert, 'I Was the Homeland Security Adviser to Trump. We're Being Hacked.' *The New York Times* (16 December 2020) <https://www.nytimes.com/2020/12/16/opinion/fireeye-solarwinds-russia-hack.html> accessed 24 March 2021; Yevgeny Vindman, 'Is the SolarWinds Cyberattack an Act of War? It Is, If the United States Says It Is' (*Lawfare*, 26 January 2021) <https://www.lawfareblog.com/solarwinds-cyberattack-act-war-it-if-united-states-says-it> accessed 24 March 2021.

[63]   Jack Goldsmith, 'Self-Delusion on the Russia Hack' (*The Dispatch*, 18 December 2020) <https://thedispatch.com/p/self-delusion-on-the-russia-hack> accessed 24 March 2021.

[64]   David E Sanger and Nicole Perlroth, 'U.S. Escalates Online Attacks on Russia's Power Grid' *The New York Times* (15 June 2019) <https://www.nytimes.com/2019/06/15/us/politics/trump-cyber-russia-grid.html> accessed 24 March 2021.

Impact of Good Corporate Practices for Security of Digital Products on Global Cyber Stability

Vladimir Radunović
DiploFoundation
Belgrade, Serbia

Jonas Grätz-Hoffmann
Federal Department of Foreign Affairs
Bern, Switzerland

Marilia Maciel
DiploFoundation
Strasbourg, France

Abstract: The exploitation of vulnerabilities in digital products and services is an essential component of sophisticated cyberattacks. Well-resourced adversaries increasingly exploit vulnerabilities for economic, political, or military gain, causing effects that destabilise cyberspace. Several multilateral and multi-stakeholder fora develop norms and principles to reduce such vulnerabilities. The main challenge lies in implementation. Under the Geneva Dialogue on Responsible Behaviour in Cyberspace[1] (Geneva Dialogue), a dozen leading global companies jointly developed a set of good corporate practices that translate high-level principles into day-to-day operations. This paper argues that these practices make cyberspace less vulnerable, and thus contribute to the implementation of global norms and principles. It further analyses key global norms and principles related to the security of digital products and services and the role of industry. It then presents the most relevant results of the ongoing work of the Geneva Dialogue, particularly good corporate practices related to security by design: threat modelling, supply chain security, development and deployment, and vulnerability processes. It discusses how these measures may reduce vulnerabilities, especially for smaller producers whose importance in the supply chain was elevated by COVID-19. It reflects on the need to turn good practices into baseline requirements to support market newcomers and regulators worldwide.

Keywords: *cybersecurity, cyber norms, vulnerability, good practices, digital products, multi-stakeholder cooperation*

---

[1] The Geneva Dialogue (https://genevadialogue.ch) is an initiative of the Swiss government and DiploFoundation. Partners of the Geneva Dialogue include Bi.Zone, Cisco, EnSign, FireEye Mandiant, Kaspersky, Huawei, Microsoft, UBS, PNG ICT Cluster, SICPA, Siemens, SwissRe, Tata Consultancy Services, VU, and Wisekey. Good corporate practices regarding the security of digital products and services, discussed in detail in this paper, have been developed through 15 group online meetings and continuous collaboration in the shared document, conducted over 7 months in 2020.

1. INTRODUCTION

Threat actors often exploit vulnerabilities in digital products, making information and communication technology (ICT) companies an initial target of their operations in order to reach their ultimate goals (Hurel and Lobato 2018). The exploitation of vulnerabilities within the supply chain of digital products by Advanced Persistent Threat (APT) actors may impose high economic costs and impact international stability. Two examples stand out. First, the NotPetya ransomware – which exploited vulnerabilities in Windows and spread through the global supply chain via a compromised update of accountancy software in Ukraine – resulted in more than US$10 billion in damages (Greenberg 2018). The US and UK governments publicly attributed the attack to the Russian military (White House 2018; UK NCSC 2018). Second, the SolarWinds hack – where a software update was compromised and was allegedly engineered by a state-sponsored APT actor (CISA 2020) – created a backdoor to about 18,000 entities (SolarWinds 2020), including US public institutions and large corporations.

The exploitation of vulnerabilities is one of the most frequent components of sophisticated cyberattacks (Uren, Hogeveen and Hanson 2018). Product security also plays a fundamental role in the development of offensive cyber capabilities, since a cyberattack is realised when the capabilities of the attackers match the possibility to exploit a vulnerability (Mladenović and Radunović 2018). Leading technical frameworks for describing sophisticated APT attacks also consider the exploitation of vulnerabilities among major components: 'Exploiting a vulnerability to execute code on a victim's system' represents the fourth phase of the Lockheed Martin Cyber Kill Chain™ (Hutchins, Cloppert and Amin 2011), while the MITRE ATT&CK framework of adversarial tactics and techniques reflects on exploiting vulnerabilities at various stages of an attack, starting with developing capabilities by 'building or acquiring solutions such as malware, exploits, and self-signed certificates' (MITRE n.d.).

Unsecured digital products allow attacks that damage global cyber stability. Therefore, states must cooperate with industry to implement international cybersecurity norms and principles (hereinafter referred to as 'norms and principles') – particularly those related to the integrity of the supply chain and the responsible reporting of vulnerabilities.

In this paper we review the related international norms and principles and discuss good corporate practices related to the security of digital products that contribute to the implementation of these norms and principles, and hence to global cyber stability.

2. THE ROLE OF THE BUSINESS SECTOR IN IMPLEMENTING CYBERSECURITY NORMS AND PRINCIPLES

Norms and principles agreed upon at the level of the UN General Assembly (UN GA) have the highest normative authority. This holds true for the report of the 2013–2015 UN Group of Governmental Experts (GGE 2015), which contains 11 voluntary norms for the responsible behaviour of states in cyberspace and has since been endorsed by the UN GA. Since then, there have not been major breakthroughs in the development of norms at the UN level. A further GGE (2016–2017) did not produce a consensus report. The debate continues in the framework of the 2018–2021 GGE (UNODA 2021). The UN Open-Ended Working Group (OEWG), which has been open to all UN member states, did produce a consensus report in March 2021. It contains an important reaffirmation of the need to implement the 11 norms agreed upon in 2015 and directs particular attention to protecting critical health infrastructure, the integrity of the supply chain and responsible reporting of vulnerabilities (UN GA 2021, 5). In 2020, Russia led the process of establishing a new 2021–2025 OEWG, which should, according to its mandate (UN GA 2020), further develop the rules, norms and principles of responsible behaviour.

In this context, better implementation of existing norms and principles is essential to enhancing cyber stability. Implementation takes different approaches at different levels. Some government-led and non-government initiatives aim to clarify, fill the gaps, or strengthen compliance with the norms developed by the GGE.

In parallel, non-government-led initiatives focusing on norms and principles have also flourished in recent years. This is an important development because the UN processes remain intergovernmental and the norms developed therein are targeted at states, even if they indirectly impact other actors. Non-government initiatives, however, expand the group of actors that hold *agency* (Passoth 2012) in promoting cyber stability and assigning active responsibilities to companies, the technical community and individuals. These normative efforts aim not only to pull non-government actors to comply with norms announced by the GGE, but also to fill gaps in these norms.

Table I shows the norms developed by the 2013–2015 GGE focusing on the security of digital products and services that have been echoed by some multi-stakeholder initiatives, including: a) the Global Commission on the Stability of Cyberspace (GCSC); b) the Paris Call for Trust and Security in Cyberspace (Paris Call); and c) the Charter of Trust.

TABLE I: A COMPARISON BETWEEN NORMS FOCUSED ON THE SECURITY OF DIGITAL PRODUCTS AND SERVICES BY THE UN GGE AND MULTI-STAKEHOLDER INITIATIVES, ADAPTED FROM GROTTOLA (2020)

UN GGE (GGE 2015)	UN OEWG (UN GA 2021)	GCSC (GCSC 2019)	Paris Call (Paris Call 2018)	Charter of Trust (Charter of Trust 2018)
Protection of the integrity of the supply chain (para 13 (i))	States should 'take reasonable steps to ensure the integrity of the supply chain, including through the development of objective cooperative measures, so that end users can have confidence in the security of ICT products'.	**Avoidance of tampering:** State and non-state actors should not tamper with products and services in development and production, nor allow them to be tampered with (Norm 3). **ICT devices and botnets:** State and non-state actors should not commandeer the general public's ICT resources for use as botnets or for similar purposes (Norm 4).	**Lifecycle security:** Strengthen the security of digital processes, products and services throughout their lifecycle and supply chain (Principle 6).	**Responsibility throughout the supply chain:** Ensure confidentiality, authenticity, integrity and availability by setting baseline standards (Principle 2). **Security by default:** Adopt the highest appropriate level of security and data protection and ensure that it is preconfigured into the design of products, processes, technologies, etc. (Principle 3).
Sharing vulnerability knowledge (para 13 (j))	States should 'encourage the responsible reporting of vulnerabilities'.	**Vulnerability equity process:** States should create transparent frameworks to assess whether and when to disclose not publicly known vulnerabilities, with the default presumption in favour of disclosure (Norm 5). **Reduce and mitigate significant vulnerabilities:** Developers and producers of products and services on which cyber stability depends should (1) prioritise security and stability, (2) take reasonable steps to ensure their products and services are free from significant vulnerabilities, and (3) take measures to mitigate vulnerabilities that are later discovered in a timely manner and to be transparent about the process (Norm 6).		

At the same time, norms must be rooted in practice and acted upon (Finnemore and Hollis 2016). The implementation of norms depends on shared ownership with engagement from both the private sector and civil society (Klimburg and Almeida 2019).

Industry plays a particular role, as the main driver and pace-setter of innovation (Kaufmann 2016), in creating digital products and owning most infrastructure. Since cyberattacks are typically executed remotely from different locations, global reach is one distinct advantage of industry over states when it comes to norm implementation. Hence, if global ICT and related industries implement similar good practices (e.g. vulnerability disclosure), norm implementation will also be advanced globally, rather than only nationally or regionally. Even though states are ultimately responsible for global cyber stability, other actors can make destabilising actions more costly by implementing existing norms and principles. This holds especially true for the private sector.

Industry generally shares an interest in having a more stable global cyberspace and protecting their business model. Early arguments have pointed to the growing economic costs of cyberattacks that would drive companies towards responsible behaviour (Anderson 2001). Voluntary corporate social responsibility, based on 'a range of corporate motives, including integrated internal motives and external pressures', is particularly important in areas in which designing holistic legal instruments is difficult (Airike, Rotter and Mark-Herbert 2016, 9).

The literature has already identified a few roles for industry in the implementation of norms, such as assisting with attribution (Fairbank 2019, 394). A mapping by the Geneva Dialogue (Rizmal and Radunović 2019) outlines several roles the corporate sector assumes and advocates for: a) information sharing on best practice and vulnerabilities, b) developing corporate norms through standardisation (focused on security by design), c) ensuring end-user security by prioritising privacy, integrity and reliability in design, and d) ensuring transparency regarding products and breaches. In addition, companies contribute to the protection of critical infrastructure and thoroughly test products (Eggenschwiler 2018).

Yet there is also growing recognition that voluntary corporate social responsibility may not be enough, and that industry must do more to enhance the security of their own products in contribution to the implementation of norms (Maxwell and Barnsby 2019). Matwyshyn (2010) warned producers were not sufficiently transparent regarding the security of their products and suggested a three-layered commitment: 1) control the security of their code, 2) warn when vulnerabilities emerge and exploitations occur, and 3) provide fixes and patches. Hathaway and Savage (2012) went further to suggest

company liability, with regulations requiring a specific vulnerability disclosure process as well as an early warning requirement, among others. Because many institutions need longer than 30 days (considered the gold standard) to apply patches – if they are able to at all – they should take greater social and legal responsibility to prevent the emergence of vulnerabilities in the first place (Hathaway 2019).

Increasing expectations, coupled with cases of APT operations that exploited vulnerabilities in widespread commercial products and with various normative initiatives and global principles, have put pressure on companies to invest more in securing their digital products. Yet, it has also become clear that the cost related to patching discovered vulnerabilities (and to reputation) surpasses the cost of embedding security throughout the development lifecycle (Dougherty et al. 2018). At the same time, the community has started mapping and discussing weaknesses in the design or implementation of security architecture (particularly in software) of various producers (Santos, Tarrit and Mirakhorli 2017). All this has incentivised companies to turn (some of) their efforts to reducing vulnerabilities during the pre-market phase, instead of (only) reacting to them once the product is on the market.

3. GOOD CORPORATE PRACTICES FOR SECURITY OF DIGITAL PRODUCTS AND SERVICES

This section of the paper serves to highlight current industry approaches to enhance the security of digital products. It draws on the findings from the Geneva Dialogue (Radunović and Grätz 2020).

3.1. Security by Design and Related Concepts

The concept of *security by design* has emerged in relation to software, hardware, services, and system integration. Geneva Dialogue partners defined it as 'designing with security in mind: addressing risks from an early stage and throughout the product development lifecycle. It may be understood as designing with security controls from the beginning' (Radunović and Grätz 2020, 5). Importantly, companies understand this as a comprehensive process that considers engineering, security, business, and human resources aspects, and involves engineers, security professionals, and C-level management.

Further, industry partners outline the *security development lifecycle* (SDL) as the most common practical model of implementing security by design. It requires producers to model security risks, driving timely decisions about reducing risk throughout the development lifecycle. SDL is particularly applied in software development but is increasingly being adapted to cloud services and internet of things (IoT) devices.

Finally, the concept of the *trustworthiness* of products relates to 'the rigorous application of design principles and concepts within a disciplined and structured set of processes that provides the necessary evidence and transparency to support risk-informed decision making and trades' (Ross, McEvilley and Carrier Oren 2016). It can be understood more broadly as a fresh perspective on SDL, which also considers non-technical issues such as internal processes and reputation (Buchheit et al. 2020), thereby relating to the trustworthiness of producers and their internal processes, rather than just products.

3.2. Good Corporate Practices

After in-depth discussions on good practices, industry partners of the Geneva Dialogue distinguished several main elements of security by design: threat modelling, supply chain and third-party security, secure development deployment, and vulnerability processes and support. In addition, they recognised the need to adjust the corporate mindset and internal processes to the security by design approach as a cross-cutting element. These elements apply across industries: software, hardware and devices, online services, and integrated systems.

3.2.1. Threat Modelling

Threat modelling is 'an engineering technique to identify possible threats, attacks, vulnerable areas, and countermeasures that could affect the product or the related environment' (Radunović and Grätz 2020, 8), which should be conducted throughout the product lifecycle and involve different departments of the company – from developers and cybersecurity specialists to senior management. Threat models depend on specific customers and the way products are implemented and used; therefore, direct cooperation with customers is recommended when possible.

Steps for performing threat modelling include: (a) identifying assets, (b) defining security requirements, (c) creating a diagram of the system, (d) identifying and analysing threats, (e) performing risk management and prioritisation, (f) mitigating threats and identifying fixes, and (g) validating mitigation (Cisco n.d.; Microsoft n.d.). In industry environments, Geneva Dialogue partners noted it is necessary to look into the system as a whole rather than focusing only on its components.

3.2.2. Supply Chain and Third-party Security

Producers commonly integrate third-party components (TPC) – both proprietary and open source – into their digital products. It is crucial that companies 'offer updates, upgrades, and patches throughout a reasonable lifecycle for their products, systems, and services via a secure update mechanism' to ensure a secure supply chain (Charter of Trust 2020a, 2). Geneva Dialogue partner practices underline the importance of a risk-based approach for digital supply chains based on three components: 1) baseline

requirements, which should also include transparency of TPC and may be an integral part of contracts, 2) supplier criticality, including defining different requirements and compliance modalities (from self-declaration and self-assessment to external audits) for various types of TPC suppliers depending on their level of criticality, and 3) verification, including establishing an internal supply chain risk management team.

Companies should create and maintain an inventory of TPC by developing a product bill of materials (BoM), creating tools for scanning and decomposition to inspect source code and images, or issuing unique IDs for hardware components. Companies should also devise a plan for when new vulnerabilities are discovered and notify suppliers about discovered TPC vulnerabilities. It is particularly important to monitor TPC that have reached their end-of-life (EoL), and thus are left without support. Suppliers, on their side, should monitor disposing of their product by EoL. Finally, producer transparency regarding the development process is crucial, and may be enhanced through transparency centres, even though the effects may be limited in cases where customers have limited knowledge of the product in question or limited resources to thoroughly check security.

3.2.3. Secure Development and Deployment

Security needs to be embedded in product development, building and testing, releasing and deployment, and validation and compliance. Security rules and checks in automated continuous integration and continuous delivery software pipelines include responsible coding, scanning source codes for vulnerabilities, dynamic analysis of code, checking dependencies for vulnerabilities, and unit tests with security checks. Particular attention must be paid to the build environment to prevent unauthorised changes, as was the case with the compromise of the SolarWinds build (CrowdStrike Intelligence Team 2021). Companies should use vetted common modules and libraries that focus on secure communications, coding and information storage.

When it comes to software, testing for vulnerabilities and validation involves static and dynamic testing, vulnerability assessment, fuzzing, penetration testing, protocol robustness testing and web application scanning. While third parties may be involved in conducting specific tests (e.g. bug-bounty programmes), a third-party audit of product and update development processes is equally important. In the case of integrated systems, testing is required for the overall configuration in addition to each of the components.

3.2.4. Vulnerability Processes and Support

Companies also set up processes to react to discovered and reported vulnerabilities by developing and distributing fixes and supporting customers. This goes hand in hand with regulatory efforts in establishing responsible vulnerability disclosure policies,

as called for by global norms. Geneva Dialogue industry partners have suggested the following elements of the process with explanations below:

- Vulnerability management: Producer 'practices and security controls to ensure products are running with the latest security updates (...) including monitoring and mitigating the effects of vulnerabilities in TPC used' (Radunović and Grätz 2020, 15). A dedicated internal product security team – often dubbed Product Security Incident Response Team (PSIRT) – should be established with a clear protocol for security servicing and plans for reacting to vulnerabilities, serving as a contact point, working in close cooperation with development, security and other teams, and issuing public security advisories.
- Vulnerability handling: Analysis of a vulnerability that is discovered or reported to the vendor, and the required remediation (i.e. developing a fix or update).
- Vulnerability disclosure: 'Overarching term for the process of sharing vulnerability information between relevant stakeholders' (Radunović and Grätz 2020, 16), related to the element below.
- Vulnerability reporting: Third-party reporting to a producer about the vulnerabilities discovered in a producer's product.
- Coordinated vulnerability disclosure: 'Coordinated information sharing and mitigation efforts about reported vulnerabilities, with producers, researchers, and other interested stakeholders' (Radunović and Grätz 2020, 17). The term *responsible vulnerability disclosure* is sometimes used instead to emphasise ethical aspects, implying a proactive investment by either party in ensuring the end goal of minimum user risk.

Importantly, this understanding of vulnerability reporting, management, and coordinated disclosure is in line with the definitions by one of the lead authorities, the Carnegie Mellon University CERT (Householder et al. 2017), while the latter is in line with the ISO/IEC 29147:2014 standard (ISO 2014).

There is a particular challenge related to the deployment of updates, since some customers may miss information about vulnerabilities and fixes and others may lack the capacity to apply them, while certain critical and complex sectors may risk their regular operations if they deploy the patch. While more research on assessing the effectiveness of patching processes is needed, it is essential that companies put more focus on preventing vulnerabilities in the first place.

3.2.5. Adjusting the Mindset and Internal Processes
Secure design demands companies to establish the right mindset throughout an

organisation; understanding security is everyone's task (Charter of Trust 2018, Principle 1). This requires 'ensuring that the organisation's people, processes, and technology are prepared to perform secure software development at the organisation level' (Dodson, Souppaya and Scarfone 2020, 4). Organisational setup should bring security and developer teams closer and enable different departments – including C-level management – to be involved throughout the product design lifecycle.

Continuous training throughout a company is essential, especially among engineers that implement security features during the design phase in cooperation with security teams. It should involve multiple teams and be practical and interactive (including games and realistic simulations). In addition, training for customers and third parties should also be provided where possible 'to help government organisations, academia, and other companies to develop skills and knowledge for product security evaluation' and 'allow them to benefit from the transparency on the product security and vulnerability related policies' (Radunović and Grätz 2020, 21).

4. ADVANCING THE IMPLEMENTATION OF GOOD CORPORATE PRACTICES TO ENHANCE CYBER STABILITY

If implemented consistently, good practices among large companies – particularly those whose products are widely used across various sectors and infrastructure – will have a wider positive impact. This has been underlined by recent major attacks enabled by design flaws, as described above. Hence, implementing the above-mentioned good practices will have a positive impact on cyber stability. This underscores the urgency to ensure that most, if not all, producers introduce security by design practices.

High complexity, market failures, unclear responsibilities, and lack of national and global cooperation are inhibiting greater security of digital products (OECD 2021). In the following section, we will discuss how increasing interdependence, regulatory action and globally agreed baseline requirements can address some of these issues, thereby contributing to the broader and more rigorous adoption of security by design practices.

4.1 Increasing Interdependence

The increasing interdependence of digital products and services (a supply chain issue) and the increasing level of criticality of ordinary services due to COVID-19 (pandemic-driven digitalisation) have enhanced vulnerabilities. The emerging IoT environment adds to this by integrating physical systems with the digital world (Carruthers 2016), allowing cyberattacks to generate even more far-reaching physical impacts.

4.1.1. Supply Chain

Digital products and services increasingly rely on TPC. This trend may be more intuitive for products such as hardware, where different manufacturers specialise to produce different components, as well as with integrated systems. It is also a trend in software development: open-source software (OSS) and off-the-shelf components have a clear advantage over in-house software (Badampudi, Wohlin and Petersen 2016). Geneva Dialogue partners warn about a risk with OSS as TPC, and examples like the Ripple20 and Amnesia:33 reports about vulnerabilities impacting the medical, transportation, energy, and retail industries are illustrative (Kol and Oberman 2020; dos Santos et al. 2020). It is therefore important that various producers, including open-source communities, embrace the elements of security by design discussed above to reduce the risks from TPC. This would contribute to a more secure supply chain – a goal set by the 2015 GGE Report (art. 13(i)), the Paris Call (Principle 6), and the Charter of Trust (Principle 2), among others.

At the same time, national security considerations play an important role in supply chain security. The development of state-sponsored attacks that exploit vulnerabilities has contributed to increased digital security risk (OECD 2021, 25). There is a risk of states influencing suppliers to embed hidden functions or weaknesses into digital products, thus making the supply chain vulnerable. A report by the UK government warns of the significant access that some states have to supply chains, which may lead to espionage and disruptive or destructive operations (UK 2019, 23). The EU invites supply chain risk assessments to also take into account non-technical factors by assessing suppliers based on inter alia the likelihood of interference from a non-EU country, the degree of control over its own supply chain, and the prioritisation of security practices (NIS Cooperation Group 2019, 22).

The increasing attention towards supply chain risks may incentivise industry to manage those risks more proactively, with broader implications for the adoption of good practices by small and medium-sized enterprises and start-ups.

4.1.2. Pandemic-driven Digitalisation

The pandemic has accelerated the overall digitalisation of society to unforeseen levels. According to McKinsey & Company (2020), companies have accelerated the digitalisation of their customer and supply chain by three to four years, while the share of digital or digitally-enabled products in their portfolios have been accelerated by seven years. Almost overnight, some ordinary services have become essential in society's 'new normal'. E-commerce, for instance, has allowed continued business cooperation (OECD 2020).

Most of these services were never conceived with security as a priority: Producers, often smaller enterprises and even start-ups, have used limited resources to focus on functionality and affordability as drivers in market competition. Their underdeveloped internal organisational culture and structure – with issues like financial and human resources – limit efforts related to security (Lavallée and Robillard 2015). There is, therefore, a need to ensure producers that may become more critical in certain circumstances embrace security by design.

It is important to underline that producers are not only IT companies. Various sectors, such as finance, health, automobiles and energy, are becoming digitalised and initiating their own digital services. Public institutions and local municipalities are also developing their own e-services – many of which have proven essential in times of crisis like that of COVID-19.

4.2. Regulations and Standards

Greater application of standards and regulatory action are also ways of enhancing the implementation of best practices. Standards related to software and device security confirm the relevance of the practices discussed in this section – yet they often do not match entirely. The Secure Software Development Framework by the US National Institute of Standards and Technology (NIST) (Dodson, Souppaya and Scarfone 2020) incorporates these practices under the framing of well-secured software and responses to vulnerabilities, which are elaborated in greater detail and with emphasis on organisational processes. In the manufacturing of hardware, software and firmware for products used in industrial systems, the discussed practices match the SDL requirements of the IEC 62443-4 standard (IEC 2019): secure implementation and coding, verification and validation, patch management and product EoL. To minimise the risk from the misuse of IoT devices, such as in botnets, the ETSI 303 645 standard (ETSI 2020) matches the discussed practices by defining baseline requirements for IoT devices: managing reports of vulnerabilities, software validation and maintenance, and security by default elements. However, it fails to directly reference threat modelling and TPC review.

Emerging regulatory frameworks also reflect on the discussed practices directly. The IoT Cybersecurity Labelling Scheme of Singapore (CSA 2020) incorporates the baseline requirements (in Tier 1) of the ETSI 303 645 standard and strengthens them with requirements (in Tier 2) for threat modelling based on the Infocom Media Development Authority of Singapore IoT Cyber Security Guide (IMDA 2020, 7) and (in Tier 3 and 4) for software testing on common errors and known TPC vulnerabilities, lists of all software components, and penetration testing (CSA 2020, 11). According to the Cybersecurity Act (EU 2019, Art. 54–55), the EU cybersecurity certification scheme shall include vulnerability disclosure policies, contact points, and a public

list of advisories. A broad range of requirements within the EU candidate certification scheme for cloud services focuses on the security of organisation and processes but will also include supply chain security, secure development environments, identification of vulnerabilities, directory and risk assessment of suppliers, controlling and monitoring third parties, and incident management (ENISA 2020, 132–144). In terms of critical sectors, the lead principles and practices for medical device cybersecurity by the International Medical Device Regulators Forum (IMDRF 2020) clearly match the main discussed practices, including threat modelling, security testing, software BoM, vulnerability disclosure, scoring, patching and support (even for legacy medical devices); the principles also add security requirements, architecture design and information sharing.

4.3. From Good Practices to Common Baseline Security Requirements

Good practices form a useful guide on how to approach security by design. Many producers, however – particularly those with limited resources and awareness – may lack incentives to invest in security by design or find it difficult to implement good practices and existing standards. To make a broader range of industries aware of and ready to embrace security by design, good practices should be used to shape the regulatory environment and assist producers in embracing the basics first.

Developing a global framework with baseline security requirements that are 'common for all digital suppliers and define the fundamentals that a supplier must address in order to ensure the cybersecurity foundations for their product/service' (Charter of Trust 2020a, 2) would be important in supporting the implementation of related norms and principles. Such baselines would also assist regulators in developing an environment based on corporate practice that is harmonised across jurisdictions.

Common baseline requirements need to account for several elements:

- Good corporate practices and requirements (e.g. by Charter of Trust (2020b));
- Regulatory instruments and requirements (e.g. labelling and certification schemes);
- Guidelines and principles of multilateral and multi-stakeholder organisations and fora (e.g. the work of the OECD and the Paris Call);
- International standards related to the security of digital products and services (e.g. International Organization for Standardization);
- Global agreements, norms and principles (e.g. GGE).

The Geneva Dialogue output document suggests that 'as the first step, a small set of very limited and universally applicable prescriptive requirements are defined'

(Radunović and Grätz 2020, 22). Particular models and challenges in developing and implementing baseline requirements should be further studied.

5. CONCLUSION

Sophisticated threat actors challenge the stability of cyberspace by exploiting vulnerabilities in digital products and services. At the level of the UN, states have endorsed norms agreed upon by the GGE in 2015. Some of these norms address the security of digital products and services. Several multi-stakeholder initiatives, such as the GCSC and the Paris Call, have also advanced principles and proposed norms on this issue. The review of norms and principles related to reducing vulnerabilities provided in this paper emphasises the important role producers have in their implementation. The industry may increasingly take on this role due to increasing public expectations about their accountability, as well as growing market incentives.

This paper presents the common understanding achieved by some leading global companies developed within the framework of the Geneva Dialogue on key concepts related to the security of products and services, such as security by design, security development lifecycle and trustworthiness. Further elaboration on good corporate practices, collected and systematised through this dialogue, distinguishes the main components of security by design: threat modelling, supply chain and third-party security, secure development and deployment, vulnerability processes and support, and changes in the corporate mindset and internal processes. A clear match with the requirements set in the related standards and regulatory frameworks confirms their applicability.

Such good practices directly contribute to the implementation of the discussed norms, and thus to global cyber stability – though further study on quantifying this effect is necessary. This paper warns, however, of the urgency to ensure all other producers embrace security by design, particularly those whose services may become more critical to society in times of crisis like that of COVID-19, as well as those whose products play an important role within global supply chains. It suggests the development of common baseline requirements to support the uniform implementation of good practices, assist a broader range of producers (especially those with limited resources), and support practice-driven and globally harmonised regulatory environments. Developing common baseline requirements should consider existing good corporate practices, regulatory instruments, global guidelines, norms and principles, and international standards. Further study of particular models and the challenges of developing and implementing such baseline requirements is suggested.

REFERENCES

Airike, Peppi-Emilia, Julia P. Rotter, and Cecilia Mark-Herbert. 2016. 'Corporate Motives for Multi-Stakeholder Collaboration – Corporate Social Responsibility in the Electronics Supply Chains'. *Journal of Cleaner Production* 131 (September): 639–48. https://doi.org/10.1016/j.jclepro.2016.04.121.

Anderson, Ross. 2001. 'Why Information Security Is Hard – an Economic Perspective'. In *Seventeenth Annual Computer Security Applications Conference*, 358–65. Washington, DC: IEEE Computer Society. https://doi.org/10.1109/ACSAC.2001.991552.

Badampudi, Deepika, Claes Wohlin, and Kai Petersen. 2016. 'Software Component Decision-Making: In-House, OSS, COTS or Outsourcing – A Systematic Literature Review'. *Journal of Systems and Software* 121 (November): 105–124. https://doi.org/10.1016/j.jss.2016.07.027.

Buchheit, Marcellus, Mark Hermeling, Frederick Hirsch, Bob Martin, and Simon Rix. 2020. 'Software Trustworthiness Best Practices'. An Industrial Internet Consortium White Paper. Industrial Internet Consortium. https://www.iiconsortium.org/pdf/Software_Trustworthiness_Best_Practices_Whitepaper_2020_03_23.pdf.

Charter of Trust. 2018. 'Charter of Trust: Our 10 Principles'. Charter of Trust. 2018. https://www.charteroftrust.com/about/.

Charter of Trust. 2020a. 'Common Risk-Based Approach for the Digital Supply Chain'. CoT Principle 2-Report 1. Charter of Trust. https://www.charteroftrust.com/wp-content/uploads/2020/02/20-02-11-CoT-P2-phase-1-report.pdf.

Charter of Trust. 2020b. 'Achieving Security by Default for Products, Functionalities, and Technologies - Baseline Requirements'. CoT Principle 3-Phase 1. https://www.charteroftrust.com/wp-content/uploads/2020/05/200212-P3-Phase-1-Baseline-Requirements_FINAL.pdf.

Cisco. n.d. 'What Is Threat Modeling?' Accessed 6 January 2021. https://www.cisco.com/c/en/us/products/security/what-is-threat-modeling.html.

CrowdStrike Intelligence Team. 2021. 'SUNSPOT Malware: A Technical Analysis'. *CrowdStrike Blog*, 11 January 2021. https://www.crowdstrike.com/blog/sunspot-malware-technical-analysis/.

Cyber Security Agency of Singapore [CSA]. 2020. 'Cybersecurity Labelling Scheme'. https://www.csa.gov.sg/-/media/csa/documents/cls/cls-pub-2--scheme-specifications-v1.pdf.

Cybersecurity & Infrastructure Security Agency [CISA]. 2020. 'Alert (AA20-352A): Advanced Persistent Threat Compromise of Government Agencies, Critical Infrastructure, and Private Sector Organizations'. 17 December 2020. https://us-cert.cisa.gov/ncas/alerts/aa20-352a.

Dodson, Donna, Murugiah Souppaya, and Karen Scarfone. 2020. 'Mitigating the Risk of Software Vulnerabilities by Adopting a Secure Software Development Framework (SSDF)'. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.CSWP.04232020.

Dougherty, Chad R., Kirk Sayre, Robert Seacord, David Svoboda, and Kazuya Togashi. 2018. 'Secure Design Patterns'. Software Engineering Institute, Carnegie-Mellon University Pittsburgh. https://doi.org/10.1184/R1/6583640.V1.

Eggenschwiler, Jaqueline. 2018. 'Geneva Dialogue on Responsible Behaviour in Cyberspace: Private Sector (Framework Document)'. Geneva Dialogue on Responsible Behaviour in Cyberspace. Zurich, Switzerland: ETH Zurich. https://genevadialogue.ch/wp-content/uploads/Geneva-Dialogue-Role-of-the-Private-Sector.pdf.

European Telecommunications Standards Institute [ETSI]. 2020. 'ETSI Releases World-Leading Consumer IoT Security Standard'. 2020. https://www.etsi.org/newsroom/press-releases/1789-2020-06-etsi-releases-world-leading-consumer-iot-security-standard.

European Union [EU]. 2019. 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). *Official Journal* L151 (7 June 2019): 15–69. https://eur-lex.europa.eu/eli/reg/2019/881/oj.

European Union Agency for Cybersecurity [ENISA]. 2020. 'EUCS – Cloud Services Scheme'. Report/Study. 20 December 2020. https://www.enisa.europa.eu/publications/eucs-cloud-service-scheme.

Fairbank, Nancy. 2019. 'The state of Microsoft?: the role of corporations in international norm creation'. *Journal of Cyber Policy* 4, no. 3: 380–403.

Finnemore, Martha, and Duncan B. Hollis. 2016. 'Constructing Norms for Global Cybersecurity'. *American Journal of International Law* 110, no. 3: 425–479.

Global Commission on the Stability of Cyberspace [GCSC]. 2019. 'Advancing Cyberstability: Final Report'. Global Commission on the Stability of Cyberspace (GCSC). https://cyberstability.org/wp-content/uploads/2020/02/GCSC-Advancing-Cyberstability.pdf.

Greenberg, Andy. 2018. 'The Untold Story of NotPetya, the Most Devastating Cyberattack in History'. *Wired* (22 August 2018). https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/.

Grottola, Stefania Pia. 2020. 'Proliferation of Cyber Norms: the Limitations of Traditional Diplomacy in Discussing Cyberconflict'. Conference paper for the 15th Annual GigaNet Symposium, 2 November 2020.

Group of Governmental Experts [GGE]. 2015. '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. https://undocs.org/A/70/174.

Hathaway, Melissa E., and John E. Savage. 2012. 'Stewardship of Cyberspace: Duties for Internet Service Providers'. Canada Centre for Global Security Studies, Munk School of Global Affairs, University of Toronto. https://www.belfercenter.org/sites/default/files/files/publication/cyberdialogue2012_hathaway-savage.pdf.

Hathaway, Melissa. 2019. 'Patching Our Digital Future Is Unsustainable and Dangerous'. *CIGI Papers*, 219. https://www.cigionline.org/sites/default/files/documents/Paper%20no.219web.pdf.

Householder, Allen D., Garret Wassermann, Art Manion, and Chris King. 2017. 'The CERT Guide to Coordinated Vulnerability Disclosure'. CMU/SEI-2017-SR-022. Software Engineering Institute, Carnegie Mellon University. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=503330.

Hurel, Louise Marie, and Luisa Cruz Lobato. 2018. 'Unpacking Cyber Norms: Private Companies as Norm Entrepreneurs'. *Journal of Cyber Policy* 3, no. 1: 61–76. https://doi.org/10.1080/23738871.2018.1467942.

Hutchins, Eric M., Michael J. Cloppert, and Rohan M. Amin. 2011. 'Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains'. *Leading Issues in Information Warfare & Security Research* 1 (2011): 80, edited by Julie Ryan, 80–106. Reading: Academic Publishing International Limited.

Infocom Media Development Authority of Singapore [IMDA]. 2020. 'Internet of Things (IoT)Cyber Security Guide'. https://www.imda.gov.sg/-/media/Imda/Files/Regulation-Licensing-and-Consultations/ICT-Standards/Telecommunication-Standards/Reference-Spec/IMDA-IoT-Cyber-Security-Guide.pdf?la=en.

International Electrotechnical Commission [IEC]. 2019. 'IEC 62443-4-2:2019'. 2019. https://webstore.iec.ch/publication/34421.

International Medical Device Regulators Forum [IMDRF]. 2020. 'Principles and Practices for Medical Device Cybersecurity'. IMDRF/CYBER WG/N60FINAL:2020. http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-200318-pp-mdc-n60.pdf.

International Organization for Standardization [ISO]. 2014. 'ISO/IEC 29147:2014'. ISO. 2014. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/04/51/45170.html.

Iswaran, S. 2020. 'Opening Speech by Mr S Iswaran, Minister for Communications and Information, Minister-in-Charge of Cybersecurity, at the ASEAN Ministerial Conference on Cybersecurity 2020'. 7 October 2020. https://www.csa.gov.sg/news/speeches/asean-ministerial-conference-on-cybersecurity-2020.

Kaufmann, Christine. 2016. 'Multistakeholder Participation in Cyberspace'. *Swiss Review of International and European Law* 26, no. 2: 217–234.

Klimburg, Alexander, and Virgilio A. F. Almeida. 2019. 'Cyber Peace and Cyber Stability: Taking the Norm Road to Stability'. *IEEE Internet Computing* 23, no. 4: 61–66. https://doi.org/10.1109/MIC.2019.2926847.

Kol, Moshe, and Shlomi Oberman. 2020. 'Ripple20: CVE-2020-11896 RCECVE-2020-11898 Info Leak'. Technical Whitepaper. Ripple20. JSOF. https://www.jsof-tech.com/wp-content/uploads/2020/06/JSOF_Ripple20_Technical_Whitepaper_June20.pdf.

Lavallée, Mathieu, and Pierre N. Robillard. 2015. 'Why Good Developers Write Bad Code: An Observational Case Study of the Impacts of Organizational Factors on Software Quality'. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering* 1, 677–87. https://doi.org/10.1109/ICSE.2015.83.

Matwyshyn, Andrea M. 2010. 'Hidden Engines of Destruction: The Reasonable Expectation of Code Safety and the Duty to Warn in Digital Products'. *Florida Law Review* 62, no. 1: 109–58.

Maxwell, Paul, and Robert Barnsby. 2019. 'Insecure at any bit rate: why Ralph Nader is the true OG of the software design industry'. *Journal of Cyber Policy* 4, no. 3: 346–361.

McKinsey & Company. 2020. 'How COVID-19 Has Pushed Companies over the Technology Tipping Point—and Transformed Business Forever'. McKinsey & Company. https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/how-covid-19-has-pushed-companies-over-the-technology-tipping-point-and-transformed-business-forever.

Microsoft. n.d. 'Microsoft Security Development Lifecycle Threat Modelling'. Accessed 6 January 2021. https://www.microsoft.com/en-us/securityengineering/sdl/threatmodeling.

MITRE. n.d. 'MITRE ATT&CK'. Accessed 6 January 2021. https://attack.mitre.org/matrices/enterprise/.

Mladenović, Dragan, and Vladimir Radunović. 2018. 'Defining Offensive Cyber Capabilities'. *Briefing and Memos from the Research Advisory Group, The Hague Centre for Strategic Studies*, GCSC Issue Brief 2 (Memo 4): 91–134. https://cyberstability.org/wp-content/uploads/2018/06/GCSC-Research-Advisory-Group-Issue-Brief-2-Bratislava.pdf.

NIS Cooperation Group. 2019. 'EU Coordinated Risk Assessment of the Cybersecurity of 5G Networks'. ENISA. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62132.

Organisation for Economic Co-operation and Development [OECD]. 2020. 'E-Commerce in the Time of COVID-19'. OECD. https://read.oecd-ilibrary.org/view/?ref=137_137212-t0fjgnerdb&title=E-commerce-in-the-time-of-COVID-19.

Organisation for Economic Co-operation and Development [OECD]. 2021. 'Understanding the digital security of products: An in-depth analysis'. OECD Digital Economy Papers, No. 305, OECD Publishing, Paris, https://doi.org/10.1787/abea0b69-en.

Paris Call for Trust and Security in Cyberspace [Paris Call]. 2018. 'Paris Call: The 9 Principles'. 11 December 2018. https://pariscall.international/en/principles.

Passoth, Jan-Hendrik, Birgit Peuker, and Michael Schillmeier. 2012. 'Introduction'. In *Agency without Actors? New Approaches to Collective Action*, edited by Jan-Hendrik Passoth, Birgit Peuker, and Michael Schillmeier, 1–11. London: Routledge.

Radunović, Vladimir, and Jonas Grätz. 2020. 'Security of Digital Products and Services: Reducing Vulnerabilities and Secure Design (Industry Good Practices)'. Geneva Dialogue on Responsible Behaviour in Cyberspace. Geneva, Switzerland: DiploFoundation. https://genevadialogue.ch/goodpractices/.

Rizmal, Irina, and Vladimir Radunović. 2019. 'Baseline Study'. Geneva Dialogue for Responsible Behaviour in Cyberspace. Geneva, Switzerland. https://genevadialogue.ch/wp-content/uploads/Geneva-Dialogue-Baseline-Study.pdf.

Ross, Ron, Michael McEvilley, and Janet Carrier Oren. 2016. 'Systems Security Engineering: Considerations for a Multidisciplinary Approach in the Engineering of Trustworthy Secure Systems'. NIST SP 800-160. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-160.

Santos, Daniel dos, Stanislav Dashevskyi, Jos Wetzels, and Amine Amri. 2020. 'Amnesia:33'. Forescout. https://www.forescout.com/company/resources/amnesia33-how-tcp-ip-stacks-breed-critical-vulnerabilities-in-iot-ot-and-it-devices/.

Santos, Joanna C. S., Katy Tarrit, and Mehdi Mirakhorli. 2017. 'A Catalog of Security Architecture Weaknesses'. In *2017 IEEE International Conference on Software Architecture Workshops (ICSAW)*, 220–223. Gothenburg, Sweden: IEEE. https://doi.org/10.1109/ICSAW.2017.25.

SolarWinds. 2020. Current Report, United States Securities and Exchange Commission. 14 December 2020. https://www.sec.gov/ix?doc=/Archives/edgar/data/1739942/000162828020017451/swi-20201214.htm.

United Kingdom Government [UK]. 2019. 'UK Telecoms Supply Chain Review Report'. UK Government, Department for Digital, Culture, Media & Sport. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819469/CCS001_CCS0719559014-001_Telecoms_Security_and_Resilience_Accessible.pdf.

United Kingdom National Cyber Security Centre [UK NCSC]. 2018. 'Russian Military "Almost Certainly" Responsible for Destructive 2017 Cyber Attack'. 14 February 2018. https://www.ncsc.gov.uk/news/russian-military-almost-certainly-responsible-destructive-2017-cyber-attack.

United Nations General Assembly [UN GA]. 2020. 'Developments in the field of information and telecommunications in the context of international security'. Accessed 23 March 2021. https://undocs.org/en/A/RES/75/240.

United Nations General Assembly [UN GA]. 2021. 'Open-ended working group on developments in the field of information and telecommunications in the context of international security Final Substantive Report'. Accessed 23 March 2021. https://front.un-arm.org/wp-content/uploads/2021/03/Final-report-A-AC.290-2021-CRP.2.pdf.

United Nations Office for Disarmament Affairs [UNODA]. 2021. 'Group of Governmental Experts'. Accessed 8 March 2021. https://www.un.org/disarmament/group-of-governmental-experts/.

Uren, Thomas, Bart Hogeveen, and Fergus Hanson. 2018. 'Defining Offensive Cyber Capabilities'. *Briefing and Memos from the Research Advisory Group, The Hague Centre for Strategic Studies* GCSC Issue Brief 2 (Memo 3): 73–90. https://cyberstability.org/wp-content/uploads/2018/06/GCSC-Research-Advisory-Group-Issue-Brief-2-Bratislava.pdf.

White House. 2018. 'Statement from the White House Press Secretary'. U.S. Embassy & Consulates in Russia. 16 February 2018. https://ru.usembassy.gov/statement-white-house-press-secretary-021518/.

The Role of Artificial Intelligence in Kinetic Targeting from the Perspective of International Humanitarian Law

Anastasia Roberts[*]
Lt Col, UK Army Legal Services
Office of Legal Affairs
SHAPE
Belgium

Adrian Venables
Senior Researcher
Department of Software Science
Tallinn University of Technology
Estonia
adrian.venables@taltech.ee

Abstract: The use of artificial intelligence (AI) in kinetic targeting is an emotive issue. Human Rights Watch (HRW) is a prominent campaigner against Lethal Autonomous Weapons Systems (LAWS) and has expressed concern these systems are fundamentally at odds with the international humanitarian law (IHL) framework for armed conflict. This framework places human control over the use of lethal force at the very heart of the targeting process. HRW asserts that the ceding of human control to AI-enabled capabilities may undermine and gradually erode the IHL framework, leaving the battlespace legally ungoverned and civilians unprotected. Concerns about the military use of AI have been exacerbated by the actions and narratives of some nations that are perceived as competing in an AI arms race. However, the debate about AI has been clouded by the fact that it focuses excessively on LAWS and human control. As a result, very little consideration is given to other potentially positive uses of AI technology in targeting. These include AI's role in Intelligence, Surveillance and Reconnaissance and Information Operations. This paper seeks to present a more nuanced examination of the role of AI in kinetic targeting and how it may affect compliance with IHL. The legal, ethical and technical arguments against and in favour of the use of AI will be examined. Finally, a way forward on this complex and emotive issue is proposed that offers a means to reinforce IHL whilst accepting that advances in technology will continue.

Keywords: *international humanitarian law, artificial intelligence, armed conflict, targeting*

[*]   The lead author is a serving member of the British Army, currently working in NATO. The views expressed in this paper are those of the author alone and not of the Army, the UK Ministry of Defence or the UK Government or of NATO.

1. INTRODUCTION

There are many potential military uses for artificial intelligence (AI) enabled technology in armed conflict.[1] However, the one that arguably attracts the most attention is its use in kinetic targeting and, in particular, the employment of Lethal Autonomous Weapons Systems (LAWS). The use of LAWS is an emotive issue as demonstrated by the high-profile Human Rights Watch (HRW) coordinated Campaign to Stop Killer Robots.[2] This campaign has been calling for an outright ban on LAWS since 2012 and continues to gather momentum.

HRW's concern is that the use of LAWS may supplant the human role in the targeting process. It sees this as being fundamentally at odds with the international humanitarian law (IHL) framework for targeting in armed conflict that centres on human control over the use of lethal force.[3] HRW asserts that ceding human control to machines may undermine and gradually erode the IHL framework, leaving the battlespace legally ungoverned and civilians unprotected.[4] Unfortunately, this focus on LAWS and human control has clouded the broader debate regarding the use of autonomous AI capabilities in kinetic targeting and distracted attention from other potentially positive uses of the technology. Some states and commentators have argued that AI could, in fact, strengthen IHL compliance in armed conflicts.[5] Regrettably, these assertions are either lost in the emotion surrounding LAWS or are met with distrust and dismissed.

Fears that future conflicts will be dominated by autonomous AI technology have been exacerbated by what commentators are now referring to as an inter-state AI arms race.[6] This is being led by the United States, China, Russia and South Korea.[7] This rivalry is reflected in the narratives of the competing nations that assert that they must

---

1   'Artificial intelligence': the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages (*Oxford Reference*, 2020) <www.oxfordreference.com/view/10.1093/oi/authority.20110803095426960> accessed 4 March 2021.

2   See campaign website at <www.stopkillerrobots.org/> accessed 4 March 2021.

3   Human Rights Watch and Harvard Law School International Human Rights Clinic, 'Killer Robots and the Concept of Meaningful Human Control – Memorandum to Convention on Conventional Weapons (CCW) Delegates' (April 2016) <https://www.hrw.org/news/2016/04/11/killer-robots-and-concept-meaningful-human-control> accessed 4 March 2021.

4   Human Rights Watch and Harvard Law School International Human Rights Clinic, 'Losing Humanity: The Case against Killer Robots' (HRW, 2012) 36 <www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots> accessed 5 March 2021.

5   US Working Paper, 'Implementing International Humanitarian Law in the Use of Autonomy in Weapon Systems' (CCW/GGE.1/2019/WP.5, 2019) paras 13–15 < https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2019/gge/Documents/2019GGE.2-WP5.pdf > accessed 4 March 2021; Peter Marguelies, 'The Other Side of Autonomous Weapons: Using Artificial Intelligence to Enhance IHL Compliance' (2018) Roger Williams Univ Legal Studies Paper No. 182 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3194713> accessed 4 March 2021.

6   Matt Bartlett, 'The AI Arms Race in 2020' (*Towards Data Science*, 16 June 2020) <https://towardsdatascience.com/the-ai-arms-race-in-2020-e7f049cb69ac> accessed 29 November 2020.

7   Justin Haner and Denise Garcia, 'The Artificial Intelligence Arms Race: Trends and World Leaders in Autonomous Weapons Development' (2019) 10:3 Glob Policy <https://onlinelibrary.wiley.com/doi/full/10.1111/1758-5899.12713> accessed 29 November 2020.

develop AI technology before their adversaries do so, fuelling a sense of urgency.[8] These provocative narratives raise the concern that states will develop capability first and then deal with the legal and moral issues afterwards.

This paper seeks to present a more nuanced examination of the use of autonomous AI in kinetic targeting and how it may affect compliance with IHL. Three potential uses of AI will be reviewed against the IHL framework: LAWS, Intelligence, Surveillance and Reconnaissance activities and Information Operations. The legal, ethical and technical arguments against and in favour of using autonomous AI technology in kinetic targeting will then be examined. Finally, a way forward on this complex and emotive issue is proposed that offers a means to reinforce IHL whilst accepting that advances in technology will continue.

2. IHL FRAMEWORK FOR TARGETING

Fundamental Principles

The International Committee of the Red Cross (ICRC) describes IHL as 'a set of rules which seek, for humanitarian reasons, to limit the effects of armed conflict'.[9] This protects those who are not, or are no longer, taking part in hostilities and reduces the suffering of those who are, for example, by proscribing weapons that cause superfluous injury or unnecessary suffering.[10] IHL is founded on four fundamental principles that underpin the targeting process: necessity, humanity, distinction and proportionality.[11] It is the latter two that raise the most challenges for the use of AI technology in targeting. The principle of distinction requires a clear difference to be drawn between civilians and civilian objects and combatants and military objects. This is necessary because only the latter may be targeted, as civilians and civilian objects are protected under IHL. The principle of proportionality requires that the incidental civilian losses resulting from an attack, known generally as collateral damage, must not be excessive in relation to the expected military advantage. This requires the use of military judgement to assess and weigh the competing military and civilian impact before an attack is authorised.

Precautionary Measures

IHL's fundamental principles and detailed rules for their application in targeting, known as 'precautions in attack', are codified in Additional Protocol I to the Geneva

---

8   Edwin Mora, 'Pentagon: U.S. "in Danger" of Losing Dominance in Artificial Intelligence' (*Breitbart*, 11 December 2018) <www.breitbart.com/national-security/2018/12/11/pentagon-u-s-in-danger-of-losing-dominance-in-artificial-intelligence/> accessed 29 November 2020.

9   ICRC, 'What is International Humanitarian Law?' (ICRC, 2004) 1 <https://www.icrc.org/en/document/what-international-humanitarian-law> accessed 4 March 2021.

10   *Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I)* (adopted 8 June 1977, entered into force 7 December 1979) 1125 UNTS 3 art 35(2).

11   ibid art 35(1); art 1(2); art 48; art 51(5)(b), art 57.

Conventions of 12 August 1949 (API).[12] Not all states are party to API but many of its provisions are considered to be customary international law (CIL) and are applicable in international and non-international armed conflict. API is clear that responsibility for applying the IHL principles and precautionary measures rests with those who plan or decide on an attack. Accordingly, whilst military commanders are supported by specialist personnel to inform their decision-making, including intelligence and legal officers, they are ultimately accountable.

When planning an attack, military commanders must do everything feasible to confirm that a selected target is military and not civilian. Furthermore, feasible precautions must be taken to avoid collateral damage, which will dictate how and when an attack is conducted. If collateral damage cannot be avoided altogether, it must be weighed against the anticipated military advantage. In this, the military commander is given a 'fairly broad margin of judgement'.[13] If the military commander assesses that the collateral damage is excessive, the attack cannot proceed. All these issues must be kept under constant review before and during a military operation. If the ongoing evaluation recognises that collateral damage is or will be excessive in relation to the military advantage expected, the attack must be cancelled or suspended. API is clear that issues of distinction and proportionality are subjective and 'must above all be a question of common sense and good faith for military commanders'.[14] It is this absence of human judgement and experience which makes the concept of autonomous AI capability so difficult to reconcile with the IHL framework.

Weapon Reviews

Article 36 of API requires state parties to review new weapons, means or methods of warfare (which are undefined) to ensure their compliance with IHL. There is no consensus as to whether this specific provision has the status of CIL thereby binding non-API states. The ICRC's view is that it does.[15] Other commentators assess that CIL requires at least a legal review of new weapons and means of warfare, if not methods.[16] In any event, according to the ICRC only a limited number of states are known to conduct legal reviews of weapons.[17]

The API commentary suggests that weapons and means are synonymous and distinguishes them from methods, which are narrowly defined as referring to how

---

[12]   Protocol 1 (n 10) art 57.

[13]   ICRC, *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (Yves Sandoz and others (eds), Martinus Nijhoff Publishers, 1987) para 2210.

[14]   ibid para 2208.

[15]   Kathleen Lawand, 'A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977' (ICRC, January 2006) 4.

[16]   Jeffrey T Biller and Michael N Schmitt, 'Classification of Cyber Capabilities and Operations as Weapons, Means, or Methods of Warfare' (2019) 95 INT'L L STUD 179, 186; William Boothby (ed), *New Technologies and the Law in War and Peace* (CUP 2018) 17.

[17]   Lawand (n 15) 5.

weapons are used.[18] However, it has been argued in the context of cyber operations that the term 'methods' is broader than this. It encompasses all tactics, techniques and procedures (TTPs) for carrying out military operations involving the conduct of hostilities, not just targeting. This decouples methods and weapons.[19] The state practice of Germany and Belgium seems to support this broader assessment but, more generally, states do not appear to have addressed this issue.[20]

Defining what constitutes a method of warfare is essential to determining whether a capability that is not obviously a weapon falls within the review process. Whether autonomous AI capability used in kinetic targeting is subject to legal review is an important element in considering whether such capability can be reconciled with the IHL framework. A legal review would need to ensure that the capability is not inherently indiscriminate and that it can apply the targeting rules, as applicable to its specific function.[21]

3. POTENTIAL MILITARY USES OF AI IN TARGETING

This section will explore three potential military uses of AI in the targeting process: LAWS, Intelligence, Surveillance and Reconnaissance (ISR) activities and Information Operations (IO).

LAWS

There is no internationally recognised definition of LAWS. The UN Group of Governmental Experts on Lethal Autonomous Weapons Systems has yet to agree on the issue.[22] In this paper, we define LAWS as a weapons system that, through the use of AI technology, can independently select and use force against targets without human control. This is likely to be achieved by the application of machine learning (ML).[23] This self-learning ability distinguishes LAWS from the so-called semi-autonomous or automated weapons systems already in use with the military. These weapons systems respond in a predefined and programmed manner to certain stimuli and are generally used in narrow defensive roles such as close anti-aircraft defence systems. A recent study of the military application of AI suggests that fully autonomous weapons

---

18 API commentary (n 13) para 1957.
19 Biller (n 16) 200.
20 Vincent Boulanin and Maaike Verbruggen, 'SIPRI Compendium on Article 36 Reviews' (SIPRI Background Paper, 2017) 3, 6 <https://sipri.org/publications/2017/sipri-background-papers/sipri-compendium-article-36-reviews> accessed 4 March 2021.
21 Boothby (n 16) 139.
22 Chair, 2020 Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, 'Commonalities in National Commentaries on Guiding Principles' (UN 2020) para 5 <https://reachingcriticalwill.org/disarmament-fora/ccw/2020/laws/documents> accessed 8 March 2021.
23 'Machine learning': the capacity of a computer to learn from experience, i.e. to modify its processing on the basis of newly acquired information (*Oxford Reference*, 2020) <https://www.oxfordreference.com/view/10.1093/acref/9780195314496.001.0001/acref-9780195314496-e-1161> accessed 29 November 2020.

systems have not yet been developed. However, both China and the US have built systems that could assume this function with simple software modifications.[24]

Intelligence, Surveillance and Reconnaissance (ISR)

Surveillance is the persistent monitoring of a target. Reconnaissance is information gathering conducted to answer a specific military question. Intelligence is the final product derived from these activities, fused with other information, which is then used to support military decision-making, including targeting.[25] It is reported that ISR is one of the areas attracting the most investment in military AI and that AI will enable dramatic improvements in this area.[26]

It is anticipated that AI will enable large amounts of information from multiple data sources to be processed and synthesised more quickly and effectively.[27] Considerable advances have already been made in image processing with some automated image-recognition and object-detection capabilities that now surpass human ability.[28] Such tools will be key to positive target identification through facial recognition but also by identifying whether observed conduct is or is not hostile. For example, is the object next to an individual digging at the side of the road an IED or a drainage pipe? Similarly, facial expression analysis could help identify hostile intent such as in the case of suicide bombers.

Information Operations (IO)

IO involves the military use of information to create a desired effect on the will, understanding and capability of adversaries and other approved parties.[29] The Internet now plays a dominant role in IO, supporting more traditional influence methods such as leaflet campaigns and radio broadcasts. It is reported that AI is already able to analyse large amounts of open-source online information to understand how to influence target audiences and tailor messaging to them.[30] As AI develops, it will also be used increasingly to create influence effects by generating, for example, autonomous online agents to engage with target audiences through social media.[31] Given the increasing prevalence of AI-generated deepfakes on the Internet, AI is also likely to be used to create and disseminate disinformation.[32] Through these means, IO

---

24   Forrest E Morgan and others, 'Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World' (RAND Corporation, 2020) 61.

25   'Joint Intelligence, Surveillance and Reconnaissance' (NATO, March 2021) <https://www.nato.int/cps/en/natohq/topics_111830.htm> accessed 9 April 2021.

26   Morgan (n 24) 20.

27   Paul Scharre, 'Artificial Intelligence: Risks and Opportunities for SOF' in Zachary S Davis and others (eds), *Strategic Latency Unleashed: The Role of Technology in a Revisionist Global Order and the Implications for Special Operations Forces* (LLNL CGSR 2021).

28   Morgan (n 24) 13–14, 17.

29   NATO, Allied Joint Publication 3.10 – Allied Joint Doctrine for Information Operations (NATO 2009) para 0107 <https://info.publicintelligence.net/NATO-IO.pdf> accessed 7 March 2021.

30   Morgan (n 24) 20.

31   ibid.

32   Robert Chesney and Danielle Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 CalLRev 1753.

may be used to facilitate the kinetic targeting process. This may involve ensuring that a target is in a desired location at a particular time or that an area is clear of civilians.

Method of Warfare?

Neither IO nor ISR activities involve the use of force unless they are integral to a weapons system. However, as noted, they may facilitate the targeting process or support targeting decisions. On this basis, the issue of whether they must remain under human control to satisfy IHL is as relevant as it is for LAWS. An autonomous capability may, for example, incorrectly identify a civilian as a target or may perfidiously feign protected status under IHL to entice a target to a particular location. Without a degree of human control to identify and prevent such occurrences, violations of IHL may result. There is a danger that the possible IHL implications of these capabilities may be missed due to the focus on LAWS.

An argument could perhaps be made that AI-enabled ISR and IO capabilities are methods of warfare and should be subject to Article 36 legal review, at least for state parties to API. This is based on the broader definition of methods of warfare as TTPs for carrying out military operations involving the conduct of hostilities, rather than simply relating to how a weapon is used. Thinking beyond LAWS would allow for a clearer discussion on the scope of the legal review process.

4. THE CASE AGAINST AI

In setting out the arguments for and against the use of autonomous AI capability in targeting, three areas are examined: legal, ethical and technical.

Legal Arguments

The primary legal concern is whether autonomous AI capabilities could even be capable of compliance with the IHL framework for targeting because they lack the requisite human judgement and experience that underlie the application of the legal tests.[33]

Distinction is an increasingly complex issue at a time when adversaries are often indistinguishable from the civilian population and will habitually alternate between targetable and non-targetable status. Often the only way to make this identification on the ground is by assessing someone's activity to discern if they are directly taking part in hostilities at a particular time, rendering them targetable. This is challenging as there is no precise definition of what constitutes direct participation in hostilities.

---

[33]   HRW (n 4) 30–34.

The ICRC CIL study[34] proposes a definition that has not been accepted by all states.[35] Moreover, while the ICRC study is helpful at a doctrinal level, the situation on the ground is often informed by the operational context and intelligence picture.

It has been suggested that the ability to discern hostility requires an understanding of an individual's mental state, which in turn relies on emotional intelligence.[36] One example of this might be celebratory weapons fire, a cultural practice in many countries. Without understanding the cultural and emotional context, autonomous AI may interpret this weapons fire as hostile activity. Moreover, as there is no clear consensus on what constitutes direct participation in hostilities and noting the key variables of operational context and intelligence, it is difficult to see how an AI capability can be programmed to learn to identify it. This would apply equally to LAWS or to standalone ISR capabilities that identify and track targets.

Dual use issues are also problematic. This is the use by the adversary of a protected civilian object for hostile purposes. In these circumstances, it must be determined whether the object has lost its protection and become a legitimate military objective. This occurs when it is making 'an effective contribution to military action' and targeting it will accordingly provide 'a definite military advantage'.[37] Again, there are no clear criteria to assess this: it is a matter of the military commander's own judgement and experience.

This is also the case with proportionality. It is difficult to see how an autonomous AI capability could conduct the required balancing exercise between military advantage and collateral damage. How will it assess the military value of the target, noting that it will be different in every attack? Similarly, how will it ascribe a value to the human life or lives involved in the context of the wider operation? These are more than ethical arguments; these are issues about compliance with the legal framework. While computer modelling and simulation are now an integral part of a collateral damage estimate for targeting, the software does not make the proportionality decision. It simply informs the military commander's decision, as does the advice received from other specialist personnel, such as legal and intelligence officers. The ICRC's position is that 'preserving human control and judgement will be an essential component for ensuring legal compliance'.[38]

Related to legal compliance is the issue of legal accountability. International criminal law provides an established framework for dealing with violations of IHL by

34   Jean-Marie Henckaerts and Louise Doswald-Beck, *Customary International Humanitarian Law Volume I Rules* (ICRC, CUP 2005).
35   John B Bellinger and William J Haynes, 'A US government response to the International Committee of the Red Cross study Customary International Humanitarian Law' (2007) 89:866 IIRC 4.
36   HRW (n 4) 31.
37   Protocol 1 (n 10) art 52.
38   ICRC, 'Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach' (ICRC, 2019) 9 <www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach> accessed 7 March 2021.

individuals. As with IHL, this framework is human-centric. If a military commander deliberately orders an attack on civilians, this is a war crime and is subject to criminal prosecution. But who is accountable for such an attack carried out or decided on by an autonomous AI capability? In cases where an autonomous AI system is intentionally manipulated by humans to commit a war crime, such as its deliberate programming to target civilians, accountability is clear. This is described as the Perpetration-by-Another liability model.[39] In these circumstances LAWS are no different to any other weapons used to commit an offence. However, perhaps a more likely and more problematic scenario is the unintended malfunction of a capability, whether LAWS or an IO or ISR capability used in support of targeting.

In the Natural-Probable-Consequence liability model, if the malfunction was the natural or probable consequence of someone's conduct, and was therefore foreseeable, that person will be held criminally accountable.[40] However, this model may be too simplistic to account for what are likely to be complex situations. In the case of the developer, for example, liability will hinge upon their level of involvement in the capability development process. If the developer was not given enough detail of the likely operational environment, including use cases and the IHL framework, it is difficult to see how foreseeability could be established.

Another suggestion is to distribute criminal accountability between key stakeholders in the creation and use of the AI capability.[41] This could include the operator, military commander, programmer, manufacturers, defence personnel involved in the acquisition process, and senior politicians. However, such an approach is likely to be evidentially challenging, politically charged and protracted, and unlikely to satisfy the victims' families. It also risks distributing accountability so widely that no individual can be held responsible for a failure under the criminal standard of proof.

Finally, the Direct-Liability model holds the AI capability itself criminally accountable.[42] Even if this was legally possible, which is debatable, it is likely to offend victims' families, making a mockery of the legal framework, and does not merit further discussion.

The lack of a clear accountability framework for IHL violations by autonomous AI capabilities is a significant impediment to the use of this technology by the military. Human accountability is a cornerstone of the IHL framework for targeting in armed conflict, and any dilution of this principle will undermine that framework.

---

[39] Gabriel Hallevy, 'The Basic Models of Criminal Liability of AI Systems and Outer Circles' (11 June 2019) 1–4 <https://ssrn.com/abstract=3402527> accessed 7 March 2021.

[40] ibid 4–8.

[41] Tetyana (Tanya) Krupiy, 'Regulating a Game Changer: Using a Distributed Approach to Develop an Accountability Framework for Lethal Autonomous Weapon Systems' (2018) 50, GJIL, 45–70.

[42] Hallevy (n 39) 8–15.

Ethical Arguments

Even if it could be demonstrated that an autonomous AI capability can comply with IHL, there is still significant opposition to the use of such capability on ethical grounds. It is argued that ceding life and death decisions to machines would deprive people of their inherent dignity and result in the dehumanisation of warfare because military decision-making would be stripped of emotion.[43] It is asserted that even when it would be lawful to use force, conscience often acts as a final barrier against killing civilians.[44] The role of human emotion over and above legal compliance was demonstrated by the International Security and Assistance Force (ISAF) policy of 'courageous restraint' in Afghanistan in 2009.[45] This encouraged military personnel to refrain from the use of force, even when legally permissible, to spare the civilian population even if at a cost to themselves or other ISAF personnel. An autonomous AI capability will not have a conscience or the human emotion to instinctively know when restraint should be exercised.

Technical Arguments

Three technical issues have emerged which suggest that AI may be unable to comply with the IHL framework. The first issue is bias. While this is not a trait usually associated with machines, it has been demonstrated that ML technology can display preferences. This is thought to be caused by the data sets it is trained with if they are unrepresentative or reflect prejudice.[46] This issue could have serious implications in the targeting process. By way of example, the training data for an ISR capability might contain a disproportionate number of images of individuals with a particular ethnicity. As a result, the capability may learn that persons of this group are *prima facie* adversaries, and therefore targetable. The question of skewed training data sets may be of particular concern where AI technology is developed internally by the military, noting that the demographic of most Western militaries is predominantly white male.

The second technical issue is the linked problems of predictability and reliability. AI/ML technology is not programmed to make decisions in a particular way but rather develops its own decision-making process by analysing and modelling its training data. As a result, developers are often unable to explain how AI technology arrived at a decision because of its complex evolving internal processes. This is known as the black-box effect.[47] If the AI decision-making process is not fully understood, it is impossible to predict how it will respond in any given situation, which reduces confidence in the system. This problem is exacerbated by the fact that the very nature

---

43   HRW (n 3).
44   HRW (n 4) 37–39.
45   Joseph H Felter and Jacob N Shapiro, 'Limiting Civilian Casualties as Part of a Winning Strategy: The Case of Courageous Restraint' (2017) 146:1 AAAS 44.
46   Select Committee on Artificial Intelligence, *AI in the UK: Ready, Willing and Able?* (HL 2017-19, 100) paras 107–121.
47   ibid paras 89–94.

of ML means that training is not finite and that a capability may keep learning from external environmental factors even when deployed.[48]

This leads to the third and most important issue of explainability; that is, the ability to describe how a decision has been made. This is a key aspect of the targeting process as a military commander must be able to explain their decision-making process to demonstrate IHL compliance. This clearly links to the issue of accountability. Accordingly, the output of any autonomous AI technology must include an analysis of its decision-making process and the factors relied on. The black-box effect described above suggests that this may be technologically impossible at this time.

5. THE CASE FOR AI

The arguments against the use of autonomous AI capability by the military in targeting must be recognised. However, the potential for AI technology to also strengthen IHL compliance is often overlooked.

Legal Arguments

Targeting in armed conflict can be fast-moving and pressured, with short decision-making windows and an imperfect intelligence picture. This is why targeting decisions are judged by the standard of a reasonable military commander and are made on the known circumstances at the time and the information available. However, if AI-enabled ISR capability develops as predicted, it will result in the faster production of a more accurate intelligence picture. The reported advances in image and facial recognition and expression analysis will provide greater certainty in distinction, positive target identification and more precise collateral damage estimates. Even the use of LAWS may in fact strengthen precautions, as when using unguided or conventional munitions, a military commander has no control over an attack once it has commenced. This means that it may not be possible to stop the attack if the collateral damage estimate changes or target identification is lost. Even with modern precision-guided munitions, control in flight remains limited. In contrast, it is suggested that LAWS will be able to abort or delay an attack as soon as it identifies a change in conditions.[49] As for IO, AI-enabled capabilities could strengthen IHL by identifying and facilitating non-kinetic alternatives to kinetic targeting, which informs the consideration of the principles of necessity and proportionality. Moreover, they may aid in reducing collateral damage by providing an effective means of warning civilians of an attack or otherwise ensuring that they are out of the target area.

In terms of the concern that the use of AI is incompatible with the concept of legal accountability, it is true that international and domestic criminal law do not appear to

---

48 ICRC (n 38) 10–11.
49 Ryan Khurana, 'In Defence of Autonomous Weapons' (*The National Interest*, 14 October 2018) <https://nationalinterest.org/feature/defense-autonomous-weapons-33201> accessed 7 March 2021.

provide a readily adaptable framework. This concern could be addressed by focusing on the accountability of the state rather than individuals. However, this approach poses its own challenges as a state cannot be held directly accountable under international criminal law. The International Humanitarian Fact-Finding Commission's ability to investigate alleged violations of IHL by state parties to an armed conflict is dependent on the consent of those parties, which is also required for any disclosure of the investigation report.[50] The injured state could refer the alleged violation to the UN but this is a political rather than legal route and the UN's response will be dictated accordingly.

Another possible option is to develop the law of state responsibility to address a state's negligent deployment, in an armed conflict, of untested or inadequately tested AI capability that operates in breach of IHL. For example, directly attacking and killing civilians in violation of the principle of distinction. Under the Draft Articles on Responsibility of States for Internationally Wrongful Acts (ASR),[51] where a state seriously violates a peremptory norm, which includes the basic rules of IHL, the legal interest of the whole international community is affected. This empowers any state to invoke the responsibility of the offending state before the International Court of Justice (ICJ), not just the injured state.[52] Indeed, there is arguably an obligation on other states to do so.[53] In the context of ICJ proceedings, whether or not the state conducted a legal review in accordance with CIL or Article 36 of API may be an important feature.

Accordingly, instead of trying to formulate new accountability models to accommodate autonomous AI capability, it may be more productive to focus on strengthening existing mechanisms for holding states to account for the development and use of such capabilities. Noting that few states appear to comply with either Article 36 of API or CIL legal review obligations, this should include strengthening legal review compliance. This could include clarifying the scope of the review process in terms of methods of warfare. A clearer accountability framework may well provide a natural brake on the rapid development of autonomous AI capability and an incentive to demonstrate compliance with the legal review process.

Ethical Arguments

It has been suggested that ethical concerns about the dehumanisation of warfare ignore the fact that IHL is deliberately structured to counter rather than endorse the effects of human emotion on the battlefield.[54] Conflict is inherently fast-paced, physically

---

50   See the International Humanitarian Fact-Finding Commission's website <https://www.ihffc.org/index. asp?page=home> accessed 7 March 2021.

51   ILC, 'Report of the International Law Commission (ILC) on the Work of its Fifty-third Session' (2001) UN Doc A/56/10 29.

52   ibid ILC commentary to art 40 ASR para 5; ILC commentary to art 48 ASR paras 8–9; Marco Sassòli, 'State responsibility for violations of international humanitarian law' (2002) 84: 846 IRRC 401, 413–414.

53   ILC (n 51) art 41(1), (2).

54   William H Boothby, *Weapons and the Law of Armed Conflict* (2nd edn, OUP 2016) 343.

demanding, mentally draining and stressful. Humans are likely to experience emotions such as fear, exhaustion and anger that may adversely influence their decision-making. The loss of comrades on the battlefield may affect their judgement and increase the risk of unlawful conduct. However, by imposing rules on the conduct of warfare, IHL seeks to control the impact of these emotions. As autonomous AI capabilities will be immune to emotion and respond to events objectively in accordance with their programming, they could actually provide exactly what IHL is seeking to achieve; that is, the best possible protection for civilians and combatants.

In light of this, it has been argued that it is no longer necessary 'to cling to a human-centred approach' to IHL on the assumption that this protection is best achieved by people. The peremptory rejection of autonomous AI technology cannot be justified by arguments about human dignity when this technology offers an alternative and potentially superior means to achieve IHL's humanitarian goals.[55] It could also be argued that someone facing death in armed conflict is more likely to be concerned about the actual loss of their life rather than who or what decides to take it. In fact, given the often-remote nature of targeting, it is likely that the origin of the decision will not be clear in any event. However, the key point is accountability in the event of the unlawful taking of life and this is perhaps where ethical arguments should focus.

Technical Arguments

It could be argued that the technical concerns about AI technology are equally applicable to humans. Humans can be biased, unpredictable and unreliable and make seemingly 'illogical and impenetrable' decisions.[56] Military commanders are only held to a reasonable standard so why would we expect more from machines? Moreover, the potential technical advantages of AI are undeniable. If realised, these will improve support to the targeting process and IHL compliance.

However, the potential benefits in terms of IHL compliance will depend entirely on how autonomous AI capabilities are programmed and utilised. Careful and conscientious development practices and compliance with the legal review process are necessary to ensure that new capabilities are legally compliant and technologically protected against interference and misuse. Technological advances will enable the imposition of constraints and increasingly complex rule sets to control the behaviour of AI-based systems. In this respect, it is important that lawyers, both military and private sector, are involved in the development of autonomous AI capability as early as possible. If legal compliance issues are identified early, rule sets to control the behaviour of the capability can be incorporated into the design, becoming an integral feature of the capability, rather than an afterthought.

---

[55] Masahiro Kurosaki, 'Toward the Special Computer Law of Targeting: "Fully Autonomous" Weapons Systems and the Proportionality Test' in Claus Kreß and Robert Lawless (eds), *Necessity and Proportionality in International Peace and Security Law* (The Lieber Studies Series Book 5, 2021).

[56] Kenneth Anderson and others, 'Adapting the Law of Armed Conflict to Autonomous Weapon Systems' (2014) 90 INT'L L STUD 386, 393.

6. CONCLUSIONS AND RECOMMENDATIONS

As noted in the introduction, the purpose of this paper was to examine the role of autonomous AI capability in kinetic targeting and its potential impact on IHL compliance. The aim was to present a more balanced analysis than is often seen because of the narrow focus on LAWS. The arguments for and against the use of such capability have been presented, hopefully demonstrating that there are in fact two sides to this debate. On the one hand, it is difficult to see how autonomous AI capability can comply with the IHL rules for targeting, given the centrality of the human role. The question is also whether this should even be attempted for ethical reasons. Let us be honest, the concept of a Terminator-style machine holding human life in its hands fundamentally feels wrong. However, on the other hand, autonomous AI technology presents clear opportunities for strengthening IHL compliance. To appreciate this fact, it is necessary to look beyond 'killer robots' to the wider use of the technology.

This leads to the fact that the current polarised debate about autonomous AI capability is unhelpful. If there is to be any meaningful control over its development, a sensible and informed middle ground must be found. Calls to ban the use of autonomous AI in military systems are both unrealistic and naïve. Any such ban will push capability development into an ungoverned space with no possibility of control or debate. The reality is that capabilities are being developed now and, without safeguards, there is a real risk of technological development outpacing IHL. This reality leads us to make several recommendations to ensure the ongoing relevance of, and respect for, IHL.

First, it should be accepted that there will never be sufficient state consensus to secure a total ban on LAWS or to introduce any new legal controls on the role of AI in military systems. The desire to employ new technology to achieve an advantage on the battlefield has been a constant feature of conflict and will not change. Instead, international organisations such as the UN should focus their efforts on supporting the development of non-binding guidance on how states should apply the existing IHL framework to this complex area.

Second, states and international organisations should specifically seek to strengthen compliance with the legal review process. Development of the non-binding guidance suggested above could be a vehicle for this. This includes clarification as to when capabilities that are not obviously weapons, but do support the kinetic targeting process, should fall under the review process as methods of warfare. A great strength of the IHL framework for armed conflict is that it is inherently flexible and is designed to adapt to incorporate new technology. States, international organisations and the public must use and trust this framework or risk losing it altogether.

Third, and linked to the above, the role of lawyers in the development of AI is key. This relates to both military and civilian lawyers, noting that much capability development takes place in the private sector. Lawyers can inform a capability's development by providing the detail of the legal framework it will need to operate within. This will allow for the development of technical rules to control the capability's behaviour. Legal compliance will then be an integral part of the design, rather than an afterthought.

Fourth, states and organisations should seek to clarify the legal accountability framework. If a clear accountability framework can be identified and, if necessary, strengthened, this may provide a deterrent effect and also slow the development of autonomous AI capability to allow for the proper consideration of legal, ethical and technical issues.

Finally, the sense of urgency being ascribed to AI development by some states, fuelled by the media and references to an AI arms race, needs to be tempered. These narratives are not helping to achieve a balanced debate on this issue. States are likely to secure greater public support and trust in their AI development initiatives if they adopt a measured, rational and open approach.

Limiting Viral Spread: Automated Cyber Operations and the Principles of Distinction and Discrimination in the Grey Zone

Monica Kaminska
Postdoctoral Researcher
The Hague Program for Cyber Norms
Institute of Security and Global Affairs
Leiden University
The Hague, The Netherlands
m.k.kaminska@fgga.leidenuniv.nl

Dennis Broeders
Full Professor of Global Security and Technology
The Hague Program for Cyber Norms
Institute of Security and Global Affairs
Leiden University
The Hague, The Netherlands
d.w.j.broeders@fgga.leidenuniv.nl

Fabio Cristiano
Postdoctoral Researcher
The Hague Program for Cyber Norms
Institute of Security and Global Affairs
Leiden University
The Hague, The Netherlands
f.cristiano@fgga.leidenuniv.nl

Abstract: The fact that States resort to automated cyber operations like NotPetya, which spread virally and have indiscriminate effects, raises the question of how the use of these might be regulated. As automated operations have thus far fallen below the threshold of the use of force, the letter of international humanitarian law (IHL) does not provide such regulation. In IHL, the principles of distinction and discrimination hold that attacks should in their targeting distinguish between the civilian population and combatants, and between civilian objects and military objectives. Attacks must not be indiscriminate, and operations that might foreseeably spread to affect civilian objects are prohibited. This paper draws inspiration from the legal principles of distinction and discrimination to suggest a non-binding norm for responsible State behaviour with regard to automated operations that fall below the threshold of the use of force: the norm proposes that States should design cyber operations so as to prevent them from indiscriminately inflicting damage. The paper finds that in the case of automated

cyber operations, a distinction between the nature of the operation and the use of the operation does not make sense because the design (nature) of the malware defines the use. In order to conform with the norm, responsible States should conduct a review of cyber operations prior to their execution. Finally, as the paper illustrates with a comparative analysis of NotPetya and Stuxnet, the post-incident forensic analysis of an operation can allow third parties and victims to determine whether the operation's designer conformed with the norm. This can help set a normative benchmark by providing a basis upon which States may call out unacceptable behaviour.

1. INTRODUCTION

Automated State-led cyber operations have the potential to spread and affect systems uncontrollably. The WannaCry and NotPetya attacks of 2017 are the most pressing examples of operations that were not designed to limit harmful effects on systems, which meant that they were able to destroy data on networks supporting a wide range of services, from national healthcare to international commercial shipping. Meanwhile, existing legal frameworks, particularly international humanitarian law (IHL), are insufficient to regulate conduct with reference to attacks like WannaCry and NotPetya that fall below the threshold of the use of force.[1] In this paper, drawing inspiration from IHL, we propose a new norm against indiscriminate cyber operations below the threshold of the use of force. The norm holds that States should design cyber operations so as to prevent them from indiscriminately inflicting damage. While the norm draws inspiration from IHL, it deviates from IHL in that it does not distinguish between lawful and unlawful objects as categories. Instead, any operation that does not seek to target a malware's payload at a particular system; that is, lacks any form of distinction and discrimination, would be considered a violation of the norm.

In considering how we might borrow from the ideas of legal weapons review and targeting law in the context of regulating automated cyber operations, we find that such operations challenge the classic IHL distinction between the 'nature' and 'use' of weapons. In order to conform to the norm, we argue that responsible State actors should conduct a normative review of cyber operations at the design stage to ensure

---

[1] While the initial NotPetya attack was launched in the context of an armed conflict between Russia and Ukraine, the malware spread globally and inflicted most of its damage outside Ukraine. The operation itself fell below the threshold of the use of force as it did not cause physical injury or significant damage beyond economic and data losses. For a longer discussion on the application of international law to NotPetya see: Michael Schmitt and Jeffrey Biller, 'The NotPetya Cyber Operation as a Case Study of International Law', EJIL: Talk! (blog), 11 July 2017, https://www.ejiltalk.org/the-notpetya-cyber-operation-as-a-case-study-of-international-law/.

that the operations are designed to limit harmful effects. This recommendation stems directly from the existing recognition in the scholarly literature that cyber weapons are not 'inherently indiscriminate' and can be designed so as to accomplish the perpetrator's goals without causing significant damage beyond the intended target.[2]

The paper is divided into three sections. First, referring particularly to recent attribution statements and State contributions to the Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security (OEWG), we argue that States are starting to get worried about automated cyber attacks, which indicates the need for the development of a norm against indiscriminate sub-threshold operations. Second, we discuss why a distinction between 'nature' and 'use' does not make sense in the context of automated cyber operations and propose that responsible States should conduct a 'normative' review of the design of cyber operations prior to their launch. Third, we compare and contrast two well-known cyber operations, NotPetya and Stuxnet, to show how a post-incident analysis of an operation can reveal whether the attacker sought to limit the operation's uncontrolled harmful effects.

2. THE NEED FOR A NORM TO LIMIT AUTOMATED ATTACKS BELOW THE THRESHOLD OF AN ARMED ATTACK

NotPetya, and WannaCry before it, forced States to think about the nature and the permissibility of automated cyber attacks below the threshold of armed conflict. The financial and operational damage done, and the indiscriminate way in which the malware spread, set these attacks apart. When a number of States coordinated their attributions of the NotPetya attack to Russia, the US and the UK made references to its automated nature. The UK condemned 'its indiscriminate design' that caused it to spread beyond its primary Ukrainian targets.[3] The United States called it out in light of the ongoing conflict between Russian and Ukraine but also underlined that 'this was … a reckless and indiscriminate cyber attack that will be met with international consequences'.[4] The unofficial American condemnation was a lot harsher. Tom Bossert, President Trump's homeland security advisor, was adamant that a spoken or unspoken red line around how the United States expects fellow countries to behave on the internet had been violated: 'The United States thinks any malware that propagates recklessly, without bounds, violates every standard and expectation of proportionality

---

2   Steven M. Bellovin, Susan Landau, and Herbert S. Lin, 'Limiting the Undesired Impact of Cyber Weapons: Technical Requirements and Policy Implications', *Journal of Cybersecurity* 3, no. 1 (2017): 61.

3   Foreign and Commonwealth Office, 'Foreign Office Minister Condemns Russia for NotPetya Attacks', GOV.UK, 15 February 2018, https://www.gov.uk/government/news/foreign-office-minister-condemns-russia-for-notpetya-attacks.

4   White House, 'Statement from the Press Secretary', 15 February 2018, https://trumpwhitehouse.archives.gov/briefings-statements/statement-press-secretary-25/.

and discrimination. Truly responsible nations do not behave this way.'[5] However, given that attacks like NotPetya take place below the threshold of the use of force – or are at least not called out by States as a use of force – the principles of IHL do not apply. In other words, there is no easy resort to principles of discrimination and proportionality to judge an indiscriminate and viral attack below the threshold.

State worries about indiscriminate cyber attacks have also surfaced in the recent and ongoing rounds of the UN processes on determining responsible State behaviour in cyberspace. The OEWG wrapped up its deliberations with a report in March 2021 and the parallel process of the UN Group of Governmental Experts (UN GGE) is still yet to be finalised.[6] States can submit their contributions in writing to the OEWG and, in contrast to the closed diplomatic process of the UN GGE, have them published on the OEWG website.[7] In January 2020, Switzerland voiced its concerns: 'While the majority of cyber operations have so far been executed in a precise and targeted manner from a technical point of view, we have recently seen cases within which cyber tools were used at random and causing unintended harmful effects.'[8] Both the first[9] and the second[10] pre-drafts of the report included an unchanged reference to this problem in the threat section: 'Pursuit of increasing automation and autonomy in ICT operations was also put forward as a specific concern.' In their responses to the first draft report, States like Brazil, Ecuador and the Netherlands explicitly supported the inclusion of this concern, the latter adding that '[t]hese independently operating and developing cyber operations are, once launched, outside the control of the initiators, and therefore the adherence to the framework of responsible behaviour including

5   Cited in: Andy Greenberg, *Sandworm: A New Era of Cyberwar and the Hunt for the Kremlin's Most Dangerous Hackers* (New York: Doubleday, 2019), 244.
6   For some background on these processes see: Tim Maurer, 'A Dose of Realism: The Contestation and Politics of Cyber Norms', *Hague Journal of the Rule of Law* (2019): 1–23; Dennis Broeders and Bibi van den Berg, 'Governing Cyberspace. Behavior, Power, and Diplomacy', in *Governing Cyberspace. Behavior, Power, and Diplomacy*, eds. Dennis Broeders and Bibi van den Berg (London: Rowman and Littlefield, 2020), 1–15; Dennis Broeders (2021) 'The (im)possibilities of addressing election interference and the public core of the internet in the UN GGE and OEWG: a mid-process assessment', *Journal of Cyber Policy*, forthcoming.
7   UNODA, 'Open-Ended Working Group', accessed 25 December 2020, https://www.un.org/disarmament/open-ended-working-group/.
8   Federal Department of Foreign Affairs FDFA et al., 'Position Paper on Switzerland's Participation in the 2019-2020 UN Open-Ended Working Group on "Developments in the Field of Information and Telecommunications in the Context of International Security" and the 2019-2021 UN Group of Governmental Experts on "Advancing Responsible State Behavior in Cyberspace in the Context of International Security"', January 2020, https://unoda-web.s3.amazonaws.com/wp-content/uploads/2020/02/switzerland-position-paper-oewg-gge-final.pdf.
9   'Initial "Pre-Draft" of the Report of the OEWG on Developments in the Field of Information and Telecommunications in the Context of International Security', n.d., https://unoda-web.s3.amazonaws.com/wp-content/uploads/2020/03/200311-Pre-Draft-OEWG-ICT.pdf.
10   'Second "Pre-Draft" of the Report of the OEWG on Developments in the Field of Information and Telecommunications in the Context of International Security', n.d., https://front.un-arm.org/wp-content/uploads/2020/05/200527-oewg-ict-revised-pre-draft.pdf.

international law cannot be ensured'.[11] As the final OEWG report[12] only reflects consensus opinions, the reference to indiscriminate cyber attacks was dropped there and moved to the Chair's summary.[13] This document contains issues that were put forward by multiple states but did not achieve consensus and will be discussed further in coming iterations of the UN processes on responsible behaviour in cyberspace.

At this point two things need to be disentangled. First, there is a conflation of automation and autonomy. While these are partly overlapping concepts, we focus in this paper on automation rather than autonomy. Autonomy is most fiercely debated in the context of Lethal Autonomous Weapon Systems (LAWS), where a whole range of ethical and legal questions are raised on the issue of (the lack of) human control and computer autonomy in military weapons and systems.[14] The debate on artificial intelligence (AI) enabled cyber attacks also touches on the issue of autonomy, as AI could enable malware to react autonomously to changing circumstances and possibilities. This debate is relatively overhyped: for most attackers, AI is not needed as the available cyber automation techniques serve their purposes.[15] This paper focuses on the automated, viral quality of cyber attacks like NotPetya and the way they spread indiscriminately. Second, if States flag automation as a problem, one of the next questions is whether this can be addressed by international law or by non-binding norms (or both). As indicated above, no State has formally stated that NotPetya violated any principles of international law, let alone IHL. Even though NotPetya seemed 'most poised to burst out of the grey zone between war and peace', State reactions indicate that it did not.[16] However, there have been some efforts to develop non-binding norms to acknowledge and address the problem of automated cyber attacks. The 2018 *Paris Call for Trust and Security in Cyberspace* explicitly acknowledges the emergence of 'malicious cyber activities in peacetime' that are 'threatening or resulting in significant, indiscriminate or systemic harm to individuals

11   *The Kingdom of the Netherlands' Response to the Pre-Draft Report of the OEWG*, n.d., https://front.un-arm.org/wp-content/uploads/2020/04/kingdom-of-the-netherlands-response-pre-draft-oewg.pdf.

12   United Nations General Assembly, 'Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security Final Substantive Report' (United Nations, 10 March 2021), https://front.un-arm.org/wp-content/uploads/2021/03/Final-report-A-AC.290-2021-CRP.2.pdf.

13   United Nations General Assembly, 'Open-ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security Third substantive session 8–12 March 2021 Chair's Summary' (United Nations, 10 March 2021), https://front.un-arm.org/wp-content/uploads/2021/03/Chairs-Summary-A-AC.290-2021-CRP.3-technical-reissue.pdf.

14   Michael C. Horowitz, 'The Ethics & Morality of Robotic Warfare: Assessing the Debate over Autonomous Weapons', *American Academy of Arts & Sciences* 145, no. 4 (Fall 2016): 25–36; Kenneth Anderson and Matthew C. Waxman, 'Debating Autonomous Weapon Systems, Their Ethics, and Their Regulation under International Law', in *The Oxford Handbook of Law, Regulation and Technology* (Oxford: Oxford University Press, 2017), https://doi.org/10.1093/oxfordhb/9780199680832.001.0001; Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: W.W. Norton & Company, n.d.).

15   Ben Buchanan et al., 'Automating Cyber Attacks' (Washington, D.C.: Center for Security and Emerging Technology, November 2020).

16   Ben Buchanan, *The Hacker and the State: Cyber Attacks and the New Normal of Geopolitics* (Cambridge, MA: Harvard University Press, 2020), 302.

and critical infrastructure' and 'welcome[s] calls for their improved protection'.[17] The Global Commission on the Stability of Cyberspace covers large-scale automated attacks by asserting that 'state and non-state actors should not commandeer the general public's ICT resources for use as botnets or for similar purposes'.[18] This norm seems primarily focused on the use of botnets, but the 'similar purposes' clause might be applicable to automated attacks like NotPetya.

In this paper we argue that viral, automated attacks could be addressed by a non-binding norm for responsible State behaviour below the threshold of the use of force that draws inspiration from legal principles derived from IHL. Norms have been constructed in this way before. Some of the eleven norms in the 2015 UN GGE report are reiterations of legal principles – such as the principle of due diligence or respect for human rights law – indicating that norms and laws are perhaps more of a continuum than a strict dichotomy.[19] Inspired by the principles of distinction and discrimination in IHL, this norm would bar indiscriminate cyber operations below the threshold of the use of force. First, under IHL, the legality of a weapon (system) is among other things determined by the fact that the weapon system cannot be indiscriminate by nature. This rule refers to the 'nature of the weapon *in the uses for which it was designed* or, as some authorities have put it, its "normal" uses; i.e., the uses for which it was intended'.[20] Second, there is the matter of the indiscriminate use of the weapon, which is covered under targeting law. The principle of distinction or discrimination requires that 'a combatant, using reasonable judgment in the circumstances, distinguish between combatants and civilians, as well as between military and civilian objects'.[21] The following section will turn to the issues that the legal review of weapons poses for automated cyber operations. It will argue that the distinction between nature and use is empty for automated cyber operations and will propose a normative review to prevent the launch of indiscriminate cyber operations.

[17] 'Paris Call for Trust and Security in Cyberspace', 12 November 2018, https://www.diplomatie.gouv.fr/IMG/pdf/paris_call_cyber_cle443433-1.pdf.

[18] GCSC, 'Advancing Cyberstability. Final Report of the Global Commission on the Stability of Cyberspace', November 2019, https://cyberstability.org/report/.

[19] Liisi Adamson, 'International Law and International Cyber Norms: A Continuum?', in *Governing Cyberspace: Behaviour, Power and Diplomacy*, eds. Dennis Broeders and Bibi van den Berg (London: Rowman & Littlefield, 2020).

[20] Anderson and Waxman, 'Debating Autonomous Weapon Systems, Their Ethics, and Their Regulation under International Law', 1105.

[21] Ibid.

3. A NORMATIVE REVIEW FOR CYBER OPERATIONS?

The two-fold normative categorisation detaching indiscriminate 'use' from indiscriminate 'nature' has long been a part of the legal debate on cyber operations.[22] The *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (*Tallinn Manual 2.0*) in fact folds, in its Rule 103, the indiscriminate use and nature dichotomy into its definition of cyber weapons, understood as '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'.[23] The definition of cyber weapons is thus embedded into that of cyber attacks (Rule 92) insofar as cyber weapons are intended to execute cyber attacks.[24] In addition, through Rule 105, the *Tallinn Manual 2.0* prohibits cyber weapons that are 'inherently indiscriminate' and can be considered, fundamentally, as 'shots in the dark'. In particular, this rule defines that 'means or methods of cyber warfare are indiscriminate by nature when they cannot be: (a) directed at a specific military objective, or (b) limited in their effects as required by the law of armed conflict and consequently are of a nature to strike military objectives and civilians or civilian objects without distinction'. Separating intentional use from 'natural' capability constitutes, however, a problematic endeavour in the assessment of cyber attacks, particularly when it comes to automated operations. Malware-like and automated cyber attacks propagate and detect unpatched vulnerabilities automatically, and thus their intentionality becomes a question of pure design. In these terms, the *modus operandi* of automated malware defies the very 'nature' vs 'use' dichotomy associated with indiscriminate attacks.

The indiscriminate use of a cyber weapon has also been traditionally defined in relation to the type of harm caused (as evidenced by Rule 103 above). This is also problematic, however, because, putting aside the fact that IHL does not apply below the threshold of armed attack,[25] the rules applying to weaponry tend to govern primarily physical effects of the kind that malware seldom achieves.[26] For example,

---

[22] Herb Lin, 'Cyber Conflict and International Humanitarian Law', *International Review of the Red Cross* 94, no. 886 (Summer 2012): 515–31; Michael N. Schmitt and Sean Watts, 'The Decline of International Humanitarian Law Opinio Juris and the Law of Cyber Warfare', *Texas International Law Journal* 50 (2015): 189.

[23] Michael N. Schmitt, ed., *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge: Cambridge University Press, 2017), 452.

[24] It must be noted that the manual explicitly rules out "the destruction of data" from its definition of cyber attack, unless the destruction of data leads to physical harm. For an alternative perspective, see: Kubo Mačák, 'Military Objectives 2.0: The Case for Interpreting Computer Data as Objects under International Humanitarian Law', *Israel Law Review* 48, no. 1 (March 2015): 55–80, https://doi.org/10.1017/S0021223714000260.

[25] Specifically with reference to the legal obligation to conduct a cyber weapons review, Kudláčková et. al. find that there is no legal obligation for States to conduct a weapons review outside Article 36 of Additional Protocol I, which is not triggered below the threshold of armed conflict. See: I. Kudláčková, D. Wallace, and J. Harašta, 'Cyber Weapons Review in Situations Below the Threshold of Armed Conflict', in *2020 12th International Conference on Cyber Conflict (CyCon)*, vol. 1300, 2020, 97–112, https://doi.org/10.23919/CyCon49761.2020.9131728.

[26] See: CCDCOE, 'Scenario 10: Legal Review of Cyber Weapons', Cyber Law Toolkit, n.d., https://cyberlaw.ccdcoe.org/wiki/Scenario_10:_Legal_review_of_cyber_weapons.

when analysing NotPetya in light of this effects-based criterion for indiscriminate use, it becomes apparent that the fake ransomware, which destroyed data, hardly compares to the physical harm caused by a weapon. Above all, this indicates how a harm-based understanding of cyber weapons hinders the protection of networks below the threshold.

States have recently attempted to address this obstacle by 'softening' the harm requirement in their interpretations of how IHL, and in particular Additional Protocol I to the Geneva Conventions,[27] applies to cyberspace. For example, France officially embraced a softer requirement for a cyber operation to rise to the level of armed attack (as defined for the purpose of IHL in Article 49 of Additional Protocol I[28]): for France, it suffices that the cyber weapon disables systems to the point that they are incapacitated 'to provide the service for which they were implemented, whether temporarily or permanently, reversibly or not'.[29] By doing so, the definition of cyber weapons interestingly shifts from harm to effects.

The legal instrument tasked with assessing the conformity of new cyber weapons to IHL standards has been outlined through *Tallinn Manual 2.0*'s Rule 110, which translates the legal review of weapons (as instituted by Article 36 of Additional Protocol I[30])to the context of cyber operations: 'All states are required to ensure that the cyber means of warfare that they acquire or use, comply with the rules of the law of armed conflict that bind them.' While constituting an important tool for the safeguarding of the principles of distinction and discrimination in wartime, a standard legal review of cyber weapons – as prescribed by the *Tallinn Manual 2.0* – cannot be applied to automated cyber operations below the threshold of armed conflict.[31] Additionally, as software remains subject to frequent changes and self-remodulations, it would be impractical to provide a new standard legal review each time software is edited.[32]

With State-sponsored cyber operations primarily occurring in peacetime and with no physical harm, a different normative approach is required to assess them and to

---

[27] The Additional Protocol I to the Geneva Conventions (1977) is a fundamental document for IHL as it reaffirms and modernizes the principles of the original Geneva Conventions (1949), https://ihl-databases. icrc.org/ihl/INTRO/470.

[28] Article 49 of the Additional Protocol I states that: '"Attacks" means acts of violence against the adversary, whether in offence or in defence.'

[29] French Ministry of the Armies, 'International Law Applied to Operations in Cyberspace', September 2019, https://www.defense.gouv.fr/content/download/567648/9770527/file/international+law+applied+to+operati ons+in+cyberspace.pdf.

[30] Article 36 of the Additional Protocol I states that: '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.'

[31] Natalia Jevglevskaja, 'Weapons Review Obligation under Customary International Law', *International Law Studies* 94 (2018): 186–221.

[32] Gary Brown and Andrew Metcalf, 'Easier Said than Done: Legal Reviews of Cyber Weapons', SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, 12 February 2014), 133, https://doi. org/10.2139/ssrn.2400530.

prevent the indiscriminate infliction of damage. This article therefore proposes to take the spirit of the principles of distinction and discrimination outside the stringent legal framework of IHL as a negative norm that: 1. applies to indiscriminate cyber operations below the threshold; 2. regards the effects of cyber operations beyond 'physical harm'; 3. proposes a 'normative' review of cyber operations focusing on nature/design; and thus, 4. creates a normative benchmark for responsible State behaviour, which can be used to hold States accountable when they fail to prevent the viral spread of their cyber operations. With the distinction between indiscriminate use and indiscriminate nature blurring away in the context of automated cyber operations, a normative review of cyber operations should prioritise the assessment of their 'design'. To this end, we use a conceptualisation of a cyber operation that primarily focuses on its nature and thus constitutes 'the combination of a propagation method, exploits, and a payload designed to create destructive physical or digital effects'.[33] Envisioning nature and use as inseparable, the next section will discuss on the basis of two contrasting cyber operations how a normative review can reveal a cyber operation's discriminate or indiscriminate design.

4. SETTING THE STANDARD IN PRACTICE: COMPARING STUXNET AND NOTPETYA

Applying a norm against indiscriminate sub-threshold cyber operations would require States to conduct a review of each operation to ascertain that the design of the operation reflects the attacker's intent to limit its uncontrolled harmful effects (including the destruction of data). This section will demonstrate how the post-incident forensic analysis of an operation, including the reverse-engineering of malware, can allow third parties and victims to determine whether an attacker conformed to the norm. Such analyses are important, as their findings provide a basis on which States can call out unacceptable behaviour and thus set a normative benchmark. Using the examples of NotPetya and Stuxnet, the section will demonstrate how, once the malware had been found 'in the wild', that is, once the malware had spread among real-world computers (as opposed to test systems),[34] interested parties were able to determine whether the operations were indiscriminate in nature from the technical analysis of the malware code.

NotPetya

NotPetya, while masquerading as ransomware, in fact irreversibly encrypted every infected machine's master boot record, thus effectively destroying these computers.[35] As a result of the operation, Maersk, just one of NotPetya's many victims, reported

---

33   Trey Herr, 'PrEP: A Framework for Malware & Cyber Weapons', *Journal of Information Warfare* 13, no. 1 (2014): 87–106.

34   Trend Micro, 'In-the-Wild - Definition - Trend Micro USA', Trend Micro, accessed 23 November 2020, https://www.trendmicro.com/vinfo/us/security/definition/in-the-wild.

35   Buchanan et al., 'Automating Cyber Attacks', 9.

the loss of 49,000 laptops and 3,500 servers.[36] There are two technical characteristics of NotPetya which reveal that it did not conform to the normative principle of discrimination at the design stages. First, an analysis of what the MITRE ATT&CK Framework[37] terms the 'initial access vector' stage of the operation reveals that the attackers compromised the software update system for the M.E.Doc financial application.[38] In order to engineer this, they had first stolen the credentials of an M.E.Doc administrator to gain control of M.E.Doc's upgrade server to modify the software update so that it would include a 'backdoored' module.[39] M.E.Doc is the most popular accounting software in Ukraine, used widely by any organisation that files taxes or conducts business in the country, including multinational corporations.[40] The attackers' choice of M.E.Doc as the attack vector, therefore, already suggests that the attackers did not pay attention to distinguishing between targets. In addition, when installed by users, the malicious update allowed the attackers to collect email usernames and passwords from organisations that use M.E.Doc and their EDRPOU numbers; these numbers are unique legal entity identifiers given to every organisation that conducts business in Ukraine.[41] The fact that the attackers engineered the malware to collect the numbers is important, as it indicates that they intended for it to spread widely and wanted to identify exactly which organisation was running the backdoored M.E.Doc software.[42] We can therefore conclude from just the analysis of NotPetya's method of delivery that it was not designed to discriminate between systems in its method of delivery.

The analyses of the 'lateral movement' stage, in which the adversary moves from one system to the next within a network, and the 'impact'[43] stage, where the adversary tries to manipulate, interrupt, or destroy systems or data, reveal a second important characteristic: the malware's high level of automation and inability to distinguish between targets before installing and releasing its payload. These stages of the operation indicate that NotPetya's designers did not seek to limit in any way the malware's uncontrolled harmful effects.

---

36   Rae Ritchie, 'Maersk: Springing Back from a Catastrophic Cyber-Attack | I-CIO', I - Global Intelligence for Digital Leaders, August 2019, https://www.i-cio.com/management/insight/item/maersk-springing-back-from-a-catastrophic-cyber-attack.

37   The framework is a matrix of adversary tactics and techniques based on real-world observations, which in a post-mortem analysis of an operation helps determine the actions an attacker might have taken. See: The MITRE Corporation, 'Matrix: Enterprise | MITRE ATT&CK™', MITRE ATT&CK™, 2018, https://attack.mitre.org/matrices/enterprise/; Chris Brook, 'What Is the MITRE ATT&CK Framework?', Digital Guardian, 23 April 2020, https://digitalguardian.com/blog/what-mitre-attck-framework.

38   Mark Simos, 'Overview of Petya, a Rapid Cyberattack', Microsoft Security, 5 February 2018, https://www.microsoft.com/security/blog/2018/02/05/overview-of-petya-a-rapid-cyberattack/.

39   David Maynor et al., 'The MeDoc Connection', Cisco Talos (blog), 5 July 2017, http://blog.talosintelligence.com/2017/07/the-medoc-connection.html; Anton Cherepanov, 'Analysis of TeleBots' Cunning Backdoor', WeLiveSecurity, 4 July 2017, https://www.welivesecurity.com/2017/07/04/analysis-of-telebots-cunning-backdoor/.

40   Greenberg, *Sandworm*, 179; Maynor et al., 'The MeDoc Connection'.

41   Cherepanov, 'Analysis of TeleBots' Cunning Backdoor'.

42   Ibid.

43   The MITRE Corporation, 'Software: NotPetya | MITRE ATT&CK™', MITRE ATT&CK™, 2018, https://attack.mitre.org/software/S0368/.

In order to propagate across systems, the NotPetya malware used a number of methods.[44] The first, and most effective, was the use of a modified version of Mimikatz, a popular open-source tool used to steal user login credentials from computer memory.[45] Once it had recovered the Windows login credentials from the machine of an infected administrative user, the malware used common Windows management tools to spread itself automatically to other systems on the same network.[46] The second method used by the malware to propagate was through the use of the EternalBlue exploit tool. EternalBlue utilises the CVE-2017-0144 vulnerability in the Server Message Block (SMB) protocol[47] on unpatched Windows systems to allow attackers to remotely infect all the systems on a given network in minutes.[48] By designing the malware so that it used not only EternalBlue but also the modified version of Mimikatz, the attackers ensured that it would self-propagate even to machines that were running an updated version of Windows.[49] NotPetya was therefore designed to behave like an automated worm, spreading via trusted networks rather than the internet, which meant that it bypassed the processes put in place to prevent ransomware attacks.[50] The presence of modified Mimikatz and EternalBlue in the malware code reveals that it was not intended to discriminate between targets, but instead was designed to propagate as widely and as quickly as possible. In fact, coupling automated credential theft and re-use with vulnerability exploitation was what made NotPetya uniquely able to propagate on the widest scale in the history of cyber attacks.[51] Most crucially, however, the malware had no mechanism to distinguish between targets prior to installing its payload: once it had spread to a new host, it automatically scanned other systems for their vulnerability to the SMB exploit in order to release its payload there as well.[52] Therefore, the indiscriminate, automated propagation and installation of malware meant that the destruction wrought by NotPetya had global ramifications.

Stuxnet

Like NotPetya, Stuxnet was also a worm with the capacity to propagate automatically, but Stuxnet serves as a good example of how a technical analysis of an operation can reveal the attackers' intent to limit indiscriminate spread and destruction. In the initial

44   CISA, 'Petya Ransomware', Cybersecurity & Infrastructure Security Agency, 1 July 2017, https://us-cert. cisa.gov/ncas/alerts/TA17-181A.
45   Alexander Chiu, 'New Ransomware Variant "Nyetya" Compromises Systems Worldwide', Cisco Talos (blog), 27 June 2017, http://blog.talosintelligence.com/2017/06/worldwide-ransomware-variant.html.
46   CISA; James Maude, 'NotPetya Ransomware: Attack Analysis | BeyondTrust', BeyondTrust, 20 October 2017, https://www.beyondtrust.com/blog/entry/notpetya-ransomware-attack-analysis; Greenberg, *Sandworm*, 182.
47   The SMBv1 protocol is a network communication protocol that was developed in 1983 to enable computers on a network to share access to files, printers, and ports. See: Carly Burdova, 'What Is EternalBlue and Why Is the MS17-010 Exploit Still Relevant?', Avast, 18 June 2020, https://www.avast. com/c-eternalblue.
48   CISA.
49   Greenberg, *Sandworm*, 182.
50   NCSC, 'Russian Military "Almost Certainly" Responsible for Destructive 2017 Cyber Attack', National Cyber Security Centre, 14 February 2018, https://www.ncsc.gov.uk/news/russian-military-almost-certainly-responsible-destructive-2017-cyber-attack.
51   Simos.
52   CISA.

access stage, in order to deliver the first iteration of the Stuxnet malware into the systems of the Natanz uranium enrichment plant in Iran, which was air-gapped, the perpetrators recruited a mole to physically infect a USB flash drive with the malware, which was then plugged into the centrifuge systems at the plant.[53] Prior to delivering the malware on the USB drive, the mole had visited Natanz a number of times in order to collect detailed information on the configuration of its systems. This allowed the attackers to update the code several times before launching the operation and ensure that the malware would only deliver its payload when it found a very specific configuration of equipment and network conditions (this stage will be elaborated on later).[54] An analysis of the intrusion vector for this first version of Stuxnet reveals that it was designed as a 'precision attack': the malware was injected into only one target network, that of the Natanz facility, and was intended to spread to systems only 'within' that network.[55]

In the second iteration of the operation, to deliver a modified version of the malware, rather than using a mole, the attackers infected the systems of five unwitting external Natanz contractors.[56] It was this change in the malware's delivery, which meant that it eventually spread outside Natanz. Although the malware was designed to only propagate automatically in 'trusted networks', the infection of the contractors' systems meant that the malware spread to the contractors' other customers, most likely through removable drives. It then spread through trusted networks, which are often channelled via the internet, and ultimately ended up infecting over 100,000 computer systems globally.[57] It was at this stage that the malware 'simply went off task'.[58] However, comparing the lateral movement stages of the NotPetya and Stuxnet operations, there is one crucial difference: while the malware spread far and wide in both cases, Stuxnet did not destroy any systems that were not its intended target because it was designed to only deliver its payload to specific types of Simatic programmable logic controller (PLC) devices.[59] Having detected a Simatic PLC, Stuxnet then verified whether it was connected to a specific type of frequency converter running at 807–1,210 Hz, which was the range within which Natanz was known to run its centrifuges.[60] When Stuxnet detected these specific configurations, it released its payload, causing the PLCs to run at different speeds; when it did not, it withheld the payload.[61] Therefore, although

---

53   Kim Zetter and Huib Modderkolk, 'Revealed: How a Secret Dutch Mole Aided the U.S.-Israeli Stuxnet Cyberattack on Iran', Yahoo News, 2 September 2019, https://news.yahoo.com/revealed-how-a-secret-dutch-mole-aided-the-us-israeli-stuxnet-cyber-attack-on-iran-160026018.html.
54   Ibid.
55   Ibid.
56   Ibid.
57   Ralph Langner, 'To Kill A Centrifuge: A Technical Analysis of What Stuxnet's Creators Tried to Achieve', *The Langner Group*, November 2013, 18.
58   Kaspersky, 'Stuxnet: Victims Zero', Kaspersky Daily, 18 November 2014, https://www.kaspersky.com/blog/stuxnet-victims-zero/6775/.
59   Nicolas Falliere, Liam O. Murchu, and Eric Chien, 'W32.Stuxnet Dossier', Symantec, November 2010, https://www.wired.com/images_blogs/threatlevel/2010/11/w32_stuxnet_dossier.pdf.
60   Jon R. Lindsay, 'Stuxnet and the Limits of Cyber Warfare', *Security Studies* 22, no. 3 (2013): 383.
61   Lindsay, 383; Michael Lee, 'Stuxnet Infected Chevron, Achieved Its Objectives', ZDNet, 9 November 2012, https://www.zdnet.com/article/stuxnet-infected-chevron-achieved-its-objectives/.

Stuxnet spread in a worm-like fashion, it did not have uncontrolled harmful effects as the malware did not release the payload in systems outside Natanz.[62] For example, Chevron, the energy company, was infected by the Stuxnet malware, but its systems did not sustain any damage.[63] In fact, so precise was Stuxnet's targeting capability that Richard Clarke, a former long-term US counterterrorism chief, commented that it felt like it had been 'written by or governed by a team of Washington lawyers' to limit its collateral damage.[64] We can therefore conclude that because Stuxnet withheld its payload outside Natanz, the spread to other networks outside the Iranian nuclear plant was highly likely to have been unintentional, while the avoidance of indiscriminate harmful effects was fully intentional. Consequently, the analysis of Stuxnet's code reveals design features which indicate that, unlike NotPetya, it complied with the norm of discrimination. Table I compares and summarises the two operations.

TABLE I: SUMMARY OF FINDINGS FROM THE ANALYSIS OF NOTPETYA AND STUXNET MALWARE

	NotPetya	Stuxnet
Initial access vector	Via backdoor implanted in M.E.Doc software update known to be used widely by civilians in Ukraine. This shows that the malware was meant to enter thousands of networks.	Via an external drive inserted directly into a single target network; and via the machines of 5 external contractors known to work at Natanz. This shows the malware was meant to enter only one network.
Lateral movement	Via trusted networks using Mimikatz and EternalBlue. This shows the malware was meant to spread rapidly into every system on the thousands of networks it entered.	Via trusted networks using a number of vulnerabilities including zero days. This shows that the malware was intended to spread, but from the 'impact' stage we can determine that the designers most likely wanted the malware to spread only in the Natanz network.
Impact	Release of malicious payload regardless of environment specifications.	Release of payload only under very specific conditions.

As the two examples illustrate, the ability to reverse-engineer malware once it has been found 'in the wild' provides a basis for judging if in the design of a cyber operation the perpetrator has complied with the norm against indiscriminate operations. In particular, an operation can be judged as indiscriminate if the analysis reveals that the malware contained no mechanism for distinguishing between 'innocent' systems and its intended target prior to installing its payload. Other indications of an operation's lack

---

62   It is important to note that although the Stuxnet malware did not release its payload in non-target systems, the attackers chose not to delete the malware from non-target systems, despite most likely having the capability to do so. See: Kim Zetter, *Countdown to Zero Day: Stuxnet and the Launch of the World's First Digital Weapon*, First Edition (New York: Crown Publishers, 2014).

63   Lee, 'Stuxnet Infected Chevron'.

64   Clarke cited in: Ron Rosenbaum, 'Richard Clarke on Who Was Behind the Stuxnet Attack', *Smithsonian Magazine*, April 2012, https://www.smithsonianmag.com/history/richard-clarke-on-who-was-behind-the-stuxnet-attack-160630516/.

of attention to discrimination in the design stages are the malware's infection vector and spreading mechanisms. If the initial access vector targets thousands of networks simultaneously, it raises the likelihood that the operation will be indiscriminate. In terms of propagation, as the analysis of Stuxnet showed, the incorporation of propagation mechanisms into the malware in itself does not necessarily indicate the attacker's lack of intent to limit the operation. Instead, propagation coupled with the malware's inability to distinguish between systems in delivering its payload is what betrays the attackers' inattention to discrimination.[65]

5. CONCLUSION

This paper has argued that the legal principles of distinction and discrimination provide inspiration for a new norm that addresses automated and indiscriminate cyber attacks below the threshold of the use of force. It showed, first, that States have started to articulate a demand for such norms, as they are increasingly concerned about indiscriminate, automated cyber operations. Second, the paper argued that to ensure compliance with the norm in the context of automated cyber attacks, the IHL distinction between the nature of the capability and the use of the capability becomes meaningless, shifting the emphasis to the notion of an 'indiscriminate cyber operation'. States should focus on reviewing the design of cyber operations to ensure that they avoid indiscriminate damage. In other words, if an automated operation is in its 'nature' designed to avoid indiscriminate damage, then its 'use' will be a direct reflection of that design. Third, the paper showed that reverse engineering malware after it has been found 'in the wild', which is routinely done in the aftermath of an operation to establish the attack's source, also allows for a determination of whether a cyber operation's designer sought to limit the harmful effects of the malware to non-target systems. Such forensic analyses are important as they provide a basis upon which States may determine if the attackers conformed with the norm and thus allow them to call out unacceptable behaviour (as part of their public attribution statements, for example) and set a normative benchmark.

---

[65] It is important to note that for the purposes of assessing compliance with the norm, it is irrelevant whether an operation was intentionally indiscriminate or indiscriminate due to coding errors or unforeseen interactions between systems. Indiscriminate spread due to negligence constitutes a breach of the norm.

Epidemic? The Attack Surface of German Hospitals during the COVID-19 Pandemic

Johannes Klick
Alpha Strike Labs
Berlin, Germany
j.klick@alphastrike.io

Robert Koch
Universität der Bundeswehr
Neubiberg, Germany
robert.koch@unibw.de

Thomas Brandstetter
Limes Security/FH St. Pölten
Hagenberg, Austria
tbr@limessecurity.com

Abstract: In our paper, we analyze the attack surface of German hospitals and healthcare providers in 2020 during the COVID-19 pandemic. A primary analysis found that 32 percent of the analyzed services were vulnerable to various degrees and that 36 percent of all hospitals showed numerous vulnerabilities. Further resulting vulnerability statistics were mapped against the size of organization and hospital bed count. The analysis looked at the publicly visible attack surface utilizing a Distributed Cyber Recon System, through distributed Internet scanning, Big Data methods, and scan data of almost 1.5 TB from more than 89 different global Internet scans. From the 1,555 identified German hospitals and clinical entities, analysis of the external attack surface was conducted by looking at more than 13,000 service banners for version identification and subsequent CVE-based vulnerability identification.

Keywords: *Distributed Cyber Recon System, German hospitals, cybersecurity, vulnerabilities, attack surface, critical infrastructure*

1. INTRODUCTION

In October 2020, US-CERT issued a warning regarding the increasing ransomware activity in the healthcare sector [1]. It was common knowledge [23] that healthcare organizations were promising targets for ransomware gangs. Surprisingly, at the very beginning of the COVID-19 pandemic, several ransomware gangs actually pledged not to hit hospitals because of the ongoing scourge. The Maze and DoppelPaymer groups, for instance, said they would not target healthcare facilities and, if they accidentally hit them, would provide the decryption keys at no charge. As another example, the Netwalker operators stated they would not intentionally target hospitals; however, if accidentally hit, the hospital would still have to pay the ransom. Unfortunately, other attacker groups did not have such scruples. Ransomware incidents against hospitals skyrocketed in October 2020, most notably with the use of Ryuk ransomware against 250 U.S.-based hospitals and clinics [20]. The criticalness of the ransomware attack wave against the U.S. was demonstrated by the very rare tri-agency ransomware alert issued by the Federal Bureau of Investigation (FBI), the U.S. Department of Health and Human Services (HHS), and the Cybersecurity and Infrastructure Security Agency (CISA), and hosted by the aforementioned US-CERT.

Naturally, in an increasingly digitized and interconnected world, those issues are not limited to the United States. In Germany in 2020, an intense discussion was prompted by an incident involving the death of a patient who had to be taken to a distant hospital because the closest hospital was signed out of emergency treatment due to an ongoing ransomware attack (see, e.g., Ralston [22]). Even though the longer ride to the more distant hospital was later found not to have been a factor in the patient's death, this specific example underscores the increasing threats posed by cyber attacks, particularly in the healthcare sector.

It must be noted, however, that cybersecurity threats in the healthcare and medical sector are anything but new. On the one hand, healthcare and medical production has always been an innovative field, in which new procedures and technologies are used. On the other hand, there are known challenges – specifically, the long life cycles, or rather the long service life of products, in this area, as well as the need for time-consuming re-certifications, such as when changing or patching the software. The need for comprehensive quality control and certification, especially in the medical field, is illustrated by the example of Therac-25 and the fatal incidents involving the faulty irradiation of patients in the 1980s [14]. Although the healthcare equipment of several vendors has a higher security level nowadays, many healthcare components and systems still have numerous security issues, some of which are even critical, according to the Common Vulnerability Scoring System (CVSS), which "provides a way to capture the principal characteristics of a vulnerability and produce a

numerical score reflecting its severity" [5]. To make matters worse, the attack surface (vulnerabilities and starting points for an attack) stemming from complex healthcare networks and equipment has become increasingly challenging to defend [13].

Given this increase in cybercrime, the question arises as to what the cybersecurity situation is in the healthcare sector, which weaknesses and vulnerabilities can be identified in the healthcare system infrastructure, and what recommendations for action can be derived from these. In the very topical context of the COVID-19 pandemic, we therefore chose to examine the cyber attack surface in terms of visible vulnerabilities of hospitals and clinical providers in our home country, Germany.

The rest of the paper is structured as follows: the chapter after the introduction gives context – it describes related work and background information on ransomware attacks in the healthcare sector, as well as the overall development of this problem. Chapter 3 describes the technical infrastructure that made our analysis possible: we describe our Distributed Cyber Recon System and how we used and extended it through our analysis. In Chapter 4 our methodology for attack surface detection of hospitals and clinical providers is presented, showing how we approached this from a healthcare entity identification point of view, as well as an attack surface correlation point of view. Chapter 5 contains the data section, where we describe the results of our findings in detail both verbally and graphically. Chapter 6 summarizes the results of our analysis, as well as its shortcomings and our ideas for future work.

2. ON THE HISTORY OF RANSOMWARE

As a result of innovation and (at that time generally) low security standards, the very first piece of ransomware, surprisingly, emerged from the medical sector. In 1989 the malware "PC Cyborg," also commonly known as "AIDS Trojan" [3], was distributed to an estimated 20,000 people, including the participants of a WHO conference on AIDS. Hidden under the guise of evaluation software, the first encryption Trojan was released; it was attributed to the American biologist Dr. Joseph Popp. Interestingly enough, the damaging effects of the Trojan were stated in the user agreements and had to be accepted by the user upfront.

Though this type of malware appeared early in computer history, it took ransomware a long time to achieve "success." Ransomware variants such as "Fake Antivirus" (2001), GPCoder (2005), CRYZIP (2006), and QiaoZhaz (2007) appeared from 2001 onwards, but the attacks were still limited, mainly due to technical reasons and/or logistical obstacles to transferring the money. Some creative approaches like WinLock used SMS and phone calls to premium numbers, for example, to monetize attacks,

but a noteworthy crime breakthrough came with CryptoLocker in 2013, introducing payments via Bitcoin. While CryptoLocker was taken down in June 2014, it was the blueprint for numerous copycats, as it showed that it was possible to earn millions within a few weeks. Thus the right combination of public-key cryptography, the digital currency Bitcoin, anonymization possibilities by using the Tor network, and providing a reliable decryption opened up a new criminal business model which today accounts for a large share of the total damage in the billions per year. Fiscutean [6] gives an overview of the history and other details of ransomware.

On the basis of various technical developments and improvements, it was thus possible for criminals to implement an effective digital extortion model by means of simple cryptography, anonymous communication, and straightforward, quasi-anonymous payment options. True, there have been cases in which the data of the attacked system has been destroyed and actual recovery was never intended (for example, because no required key material was kept). However, these were exceptions and stemmed either from errors in technical implementation or simply from the attacker having other intentions than demanding ransom. The success was based on the fact that victims who choose to pay have a good chance of recovering their data; the attackers are thus motivated to enable correct decryption in order to keep their business model alive and thriving.

In the earlier days, the attackers struck out at random; the victims were often individuals and typically could pay only small ransoms. However, over time, the attackers grew more professional. They began targeting large organizations, and their attacks and ransom demands became bolder [18]. Companies active in the grey area, which sell vulnerabilities including 0days (vulnerabilities that are still unknown to the manufacturer of the product), extend this threat. An example of this is the "MedPack" of the company GLEG Ltd., which contains 0days specifically for medical software [15].

The amount of the ransom is, for obvious reasons, based on a corresponding analysis of the target. Blackmailers have also increased the pressure on the victims by threatening to publish data stolen from the company. On several occasions, they have followed through [11].

In theory, this trend can only be interrupted if no more payments are made for a long period of time. The technical prerequisite and basis for this are regular offline backups, as well as regular tests of the disaster recovery procedures, with dedicated checks on ransomware recovery.

Unfortunately, theory and practice are often worlds apart. As Goodwin and Smith [25] found, only half of all apps are fully covered by a disaster recovery strategy. In some cases, backups are not current and up-to-date or just not available due to misconfiguration, perhaps because they are not kept offline and were also encrypted. In other cases, critical aspects of the recovery process fail because they were never validated in the current environment.

Driven by the increase in ransomware attacks, companies are considering either investing in cyber insurance in order to cushion potential financial damage, or, if the situation arises, simply paying the ransom. The statistics are telling: over 40 percent of cyber insurance claims now involve ransomware [4]. Accordingly, some countries are considering banning the payment of ransoms in order to undercut the criminals' business model. The U.S. Department of the Treasury has already pointed out that ransom payments to groups or organizations on the sanctions list are punishable if they are not approved [24] by the Office of Foreign Assets Control (OFAC). "Cyber-related Sanctions" is a special section on the U.S. Department of the Treasury's website.

The difficulty of implementing this requirement is, however, evident from past cases such as the attacks on police departments in Swansea, Massachusetts [17], and in Dickson, Tennessee. These departments, infected by the ransomware CryptoWall 2.0, paid a ransom to recover their data. With this background, it is worthwhile to explore the attack surface and security posture in the healthcare sector.

Given the increase in ransomware campaigns, the outstanding importance of a functioning healthcare system – especially in the prevailing COVID-19 pandemic – and possible influencing factors through the short-term provision and integration of remote access and teleworking possibilities, we have conducted an in-depth investigation of the cyber attack surface of German hospitals based on the "Deutsche Interdisziplinäre Vereinigung für Intensiv- und Notfallmedizin" (DIVI) intensive register [9]. The DIVI register is a national registry of intensive care capacities detailing available and overall intensive-care capacities in Germany. It was created in response to the COVID-19 pandemic.

3. INTRODUCTION TO THE DISTRIBUTED CYBER RECON SYSTEM (DCS)

The previous chapters have illustrated that it is both possible and feasible to attack hospitals and medical devices. However, the question arises: just how large is the potential cyber attack surface of critical infrastructures like hospitals?

Reconnaissance and, in particular, representation of an organization in cyberspace is a major challenge. For this reason, there was a need for a novel search engine that could search the Internet (2.8 billion routed IPv4 addresses) in a few hours for a network service or services/servers with a specific vulnerability in a matter of hours, and which would also allow mapping to a specific target organization.

In our Distributed Cyber Recon System (DCS) developed specifically for various reconnaissance and analysis tasks, we can answer questions like: What is the security posture of a specific organization? What is the attack surface of an entire group of organizations? Which systems belong to which organization in the first place? Since plain Internet scan data is not sufficient, the scan data is augmented with additional information such as WHOIS data, IP prefix, Autonomous System (AS) information, certificate information, and geo-information about the IP of the system. The combination of this information in a Big Data approach enables a quite accurate representation of cyberspace.

For example, the distribution of selected system versions of a particular network service in an organization, as well as all detected Industrial Control Systems (ICS), can be displayed on a map, and systems can be organized by, e.g., specific country. IP prefix and IP ownership information can also be selected and aggregated using dynamic charts. This allows a recon analyst to get a quick overview of their own cyber infrastructures, as well as those of foreign states, organizations, and companies.

In our case, this DCS was used to analyze the publicly visible system attack surface of hospitals located in Germany. In the following passages, the methodology of our data collection, and that of the DCS, is explained in more detail.

The DCS searches the Internet globally from 1,024 different IPv4 addresses. First, TCP SYN scans are performed for 2.8 billion IPv4 addresses. For each response to this scan, the corresponding application protocol, such as HTTPS or Telnet, is scanned. Then, for each IP address, the result is enriched with owner information (Autonomous System (AS) Information), holder information (WHOIS database), Geo-IP data, BGP information, and other data sources from the Internet. For the analysis of a target entity, all data fields in the scan data are searched for the name or domain of the entity. The identified IP addresses and reachable network services are then used for further analysis. Consequently, the DCS always scans the entire Internet and only identifies the associated network areas and network services of a target entity in a post scan phase. This means the DCS database essentially holds the same type of data for any conceivable target set. In the next stage, this target information is made available in a user interface called Inspector for further advanced analyses such as vulnerability matching based on the service banner, subdomain identification, and screenshot generation.

The DCS primarily consists of the search nodes, a backend, and a frontend. The relationships between the individual components are shown in Figure 1. The frontend is used by an analyst for operation setup and data analysis. The IPv4 network ranges, protocols, ports, and scan algorithms to be scanned are defined in the frontend.

FIGURE 1: DISTRIBUTED CYBER RECON SYSTEM ARCHITECTURE
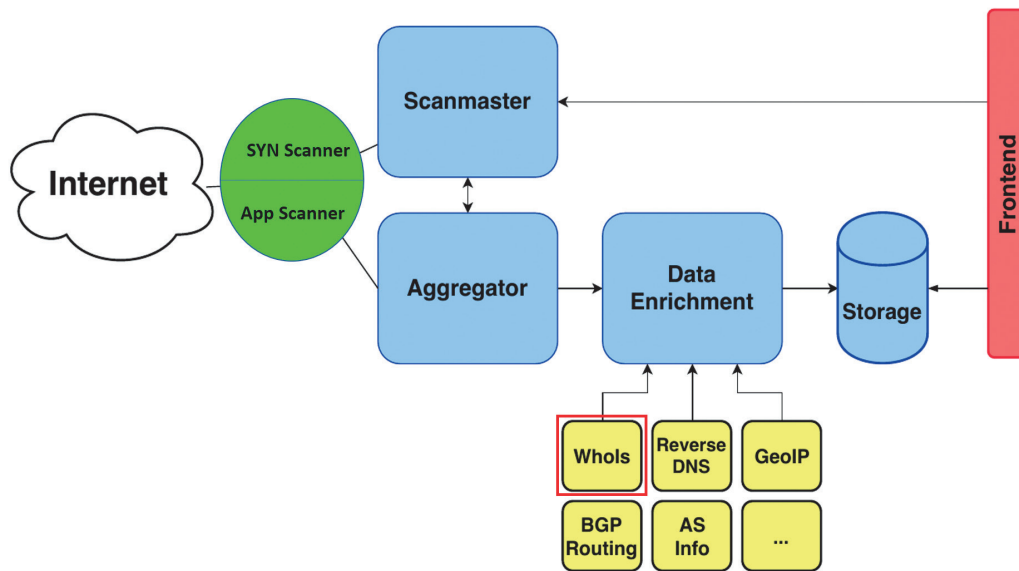
The IPv4 range to be scanned is then pseudorandomized by the scan master in accordance with the selected scan algorithm, divided into several work units and distributed to the various scan nodes. In addition, this measure helps to stay below the triggers of intrusion prevention systems, because the scan traffic is distributed to as many different target networks as possible at the same time. The scan nodes are distributed worldwide for quality and correlation reasons and have different scan bandwidths.

Our DCS enables us to scan the same target areas simultaneously from different strategically interesting locations (e.g., different countries) as site groups and to compare the results. With the help of well-chosen scan locations, potential national Geo-IP blocks can be detected and subsequently bypassed. Experience has shown that result quality can significantly improve with a globally distributed group of scan nodes, as not all destinations are visible from all parts of the Internet due to various national or regional filtering approaches. Furthermore, if a scan node fails, the scan master will automatically detect this. Subsequently, the scan master will assign the work unit of the failed node to a new search node. This ensures that all required IP addresses are always scanned, guaranteeing that the system produces consistent data.

The search nodes consist of two primary components. First of all, the SYN scanner is active, which only sends TCP SYN or UDP packets. During the sending process, the last used destination addresses are stored in a ring buffer. At the same time, the scanner is waiting for incoming TCP SYN ACK packets or UDP responses whose senders correspond to the destination addresses of the ring buffer. This prevents the search engine from being used as a DDoS amplifier. The ring buffer ensures that the search engine only responds to TCP SYN-ACK packets that it has sent out itself. Without the ring buffer, an attacker could send spoofed TCP SYN-ACK packets via IP spoofing, and the search engine would send additional application protocol level scan traffic to the spoofed IP addresses, thus using the search engine as a DDoS amplifier. Furthermore, the *search nodes use a total of more than 1,024 different IPv4 originator addresses* and can thus distribute the scan traffic. This allows the search to stay below the radar of many intrusion prevention systems and thus increases the scan data quality significantly.

As soon as a valid packet arrives from a destination address, the application scanner is started. The application scanner supports more than 60 different protocols and establishes full application connections with the goal of reading as much identification information as possible from the system. Most of the protocols we implemented ourselves; for several standard protocols, we used the Zgrab implementation [21].

After the IPv4 addresses of a work unit are processed, the scan results are sent to the aggregator. The aggregator collects all results from search nodes and checks them for consistency. Then the data is enriched with other information from open sources in JSON format.

For example, the IPv4 scan data is enriched with the INETNUM and WHOIS information from the RIRs (RIPE, ARIN, AfriNic, etc). Possible inconsistencies within the databases, such as overlapping prefixes, are resolved according to a self-developed method defined in [12]. In addition, the BGP data valid at the respective time is stored for each IPv4 address. This includes the BGP prefix annotated at the time, including all available autonomous system data. As a data source for the BGP information, the data of the Cooperative Association for Internet Data Analysis (CAIDA) [10] is merged and processed. In addition, reverse DNS records and Geo-IP information are added to each discovered active IPv4 address of the respective scan. All data is stored in a NOSQL-based ElasticSearch database, which can be duplicated as an on-premises solution for private discretionary analysis – commonly needed by, for instance, defense organizations – at any time.

For the analysis of the hospital data, a separate subfrontend called Inspector was developed, to make this complex task easier for our human analysts. The Inspector

only receives the names of the hospitals and the respective domain name as input. Subsequently, all relevant entries in the database, such as the WHOIS description field, or the common names of the collected TLS certificate information or the atomic system data, are analyzed for membership of the respective target set using self-developed advanced Big Data algorithms. In parallel, all subdomains of the added domains are searched. This is done by special best guess algorithms or by searching known public certificate databases. The Inspector had to be created, as our analysts had to take a huge list of potential healthcare target organizations into account.

The final step is about vulnerability detection: after all network services of the defined reconnaissance targets, in our case hospitals and other healthcare providers, had been identified, the system descriptions or version strings read out were compared with the National Vulnerability Database (NVD) of NIST [19]. Through this step, all potential known vulnerabilities in detected software systems are identified.

4. METHODOLOGY OF ATTACK SURFACE DETECTION OF HOSPITALS AND CLINICAL PROVIDERS

To identify the attack surface of the German hospitals, the German hospitals themselves first had to be identified. Therefore, we chose as a starting point the German DIVI registry [9], which was first established during the COVID-19 pandemic.

The DIVI Intensive Care Register records the free and occupied treatment capacities in intensive care medicine at about 1,300 acute hospitals in Germany on a daily basis. During the pandemic and beyond, the registry makes it possible to identify bottlenecks in intensive medical care, regionally and/or temporally. Thus the DIVI Intensive Care Registry creates a valuable basis for response and data-driven action control in near real-time since April 2020.

Our approach was the following: we extracted over 1,300 names of German hospitals with COVID-19 intensive care units from the DIVI Register. We then manually searched for the main website or domain of the corresponding hospital names and added them to the DIVI Registry data.

In the next processing step, the names and domain information were entered into the Inspector. The Inspector then analyzed a total of 89 different port/protocol scans. A sizable amount of data – 1,483 GB – was analyzed on a system with 1 TB Ram, 64 CPU cores, 40 TB SSD storage and 72 TB HDD storage. The total computing time of the whole system was about 16 hours.

Table I is a listing of the port/protocol combinations for which global scans for about 2.8 billion routed IPv4 addresses have been conducted.

After identification of the associated network services based on the certificate information and WHOIS and BGP/AS data, as well as the extended detection of subdomains, it was possible to collect additional information about other hospitals. For example, the cryptographic TLS certificate of Hospital A might also include the domain of another Hospital B of the same provider. Furthermore, generic search terms such as "hospital" and "clinic" were added. In addition, the results were manually searched, and any false positives were eliminated. This approach made it possible to extend the analysis of 1,300 hospitals of the DIVI registry to 1,555 hospitals.

TABLE I: SCANNED TCP AND UDP PORTS DURING ATTACK SURFACE MAPPING

http-1000	bacnet-47808	postgres-5432	openport-1025
http-5985	bigip-443	qnapvuln-8080	openport-111
http-7547	cve20205902-443	redis-6379	openport-11211
http-80	dnp3-20000	s7-102	openport-11711
http-8008	imap-143	samba-445	openport-1201
http-8080	ipmi-623	snmpv1-161	openport-135
http-8088	ipp-631	snmpv2-161	openport-139
http-8888	kibana-5601	ssh-22	openport-1433
https-1443	knx-3671	ssh-22022	openport-1521
https-443	ldap-389	ssh-2222	openport-1720
https-4433	ldapudp-389	sworionrest-17778	openport-1723
https-4434	modbus-502	telnet-23	openport-199
https-4444	mongodb-27017	telnet-2323	openport-2012
https-5986	mssqludp-1433	telnet-4786	openport-27017
https-8443	mssqludp-1434	telnet-5938	openport-3306
dnstcp-53	mysql-3306	telnet-7070	openport-3389
elastic-9200	netbios-137	upnp-1900	openport-445
eniptcp-44818	ntp-123	winrm-5984	openport-469
fox-1911	oracledb-1521	openport-993	openport-5037
ftp-21	pop3-110	openport-995	openport-5432
openport-873	openport-6379	openport-5900	openport-5555
openport-9200	openport-8009	openport-5984	openport-5601
openport-587			

5. DATA SECTION

Our analysis of the 1,555 German hospitals revealed a digital attack surface of 13,497 network services, or 8.7 network services per hospital on average. Figure 2 shows the distribution of the main service banner groups of all identified hospital network services which were identified by executing a full handshake.

Approximately 47 percent of all collected service banners are empty and thus comply with the common best-practice approach of not disclosing any software version information via service banner. This approach is very important because it makes it more difficult for attackers to identify the software used. This makes it subsequently harder for a potential attacker to determine the proper exploit/malware to use in an attack attempt. This is especially true for the use of automated attack scripts, often used by automated botnets.

We identified 1,228 hospitals and hospital operating companies that had network services that could be directly located. Approximately 300 other hospitals had no network services of their own but only those that could be assigned to joint operating companies. However, since we do not know how the networks of the joint hospital operating companies are related to the hospitals, we consider the whole operating company as a single hospital. Thus we technically analyze 1,228 hospital entities and operating companies representing up to 1,555 different hospitals. Of the 1,228 hospitals, 447 had vulnerable network services. This means that 36.4 percent of all identified hospitals and hospital operating companies have vulnerabilities.

Figures 3, 4, and 5 show the version distribution of the three most common web servers: Apache httpd, Microsoft IIS, and nginx. A well-known problem in the industrial and (to a certain degree) the healthcare sector became visible quite early in our analysis: outdated services for which end-of-support had already been announced. The most noteworthy candidates we identified included Apache httpd Version 2.2.x, which became end-of-support in December 2017, or Microsoft Internet Information

Services 6.0, which became end-of-support in June 2015. It is unclear, however, why we found those legacy services on Internet-facing systems, as the issue of patch and update difficulty typically affects mainly internal medical components, not Internet infrastructure.

FIGURE 3: VERSION DISTRIBUTION OF DETECTED APACHE WEB SERVERS, WITH ROUGHLY ONE-THIRD HAVING KNOWN VULNERABILITIES. NOTE THAT 2,092 APACHE SERVERS (68.43 PERCENT) RESULTED IN AN UNDEFINED VERSION AND ARE NOT INCLUDED

No CVEs 6.1%
Apache 2.4.29 15.2%
Apache 2.4.25 10.7%
Other Versions with CVEs 35.8%
Apache 2.4.38 9.5%
Apache 2.4.18 9.1%
Apache 2.4.10 7.4%
Apache 2.4.6 6.2%
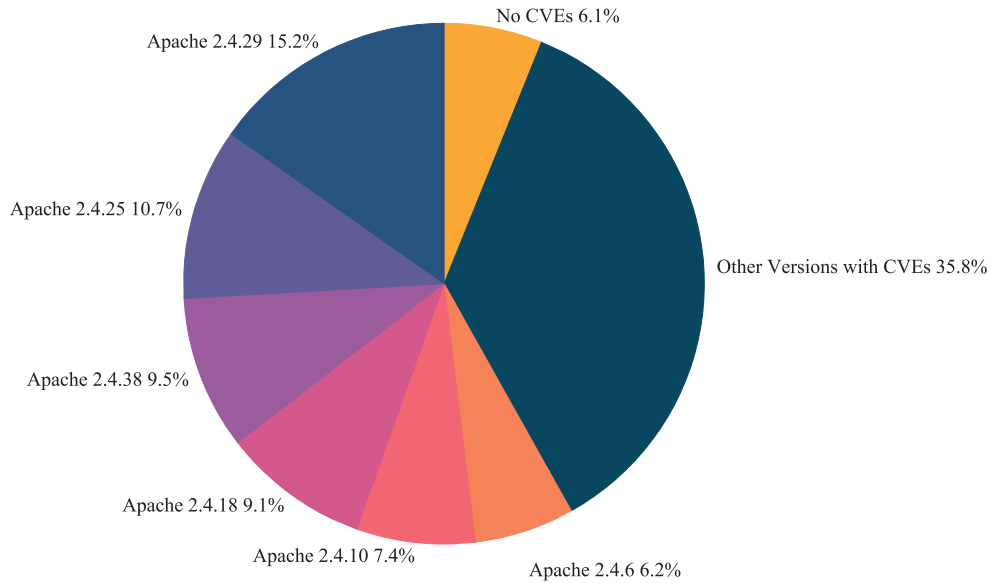
FIGURE 4: DISTRIBUTION OF DETECTED VERSIONS OF MICROSOFT INTERNET INFORMATION SERVICES (IIS) WEBSERVER, INDICATING CURRENT AS WELL AS END-OF-SUPPORT VERSIONS IN OPERATION

Microsoft−IIS undefined 0.6%
HTTPAPI/2.0 21.5%
Microsoft−IIS/8.5 31.8%
Microsoft−IIS/10.0 25.7%
Microsoft−IIS/8.0 2.1%
Microsoft−IIS/7.5 16.8%
Microsoft−IIS/6.0 1.0%
Microsoft−IIS/7.0 0.6%

Figure 6 shows the geographic location of all 1,300 hospitals of the DIVI register. This clearly shows that there is a high density of hospitals, particularly in the densely populated regions of western Germany and in the German metropolitan areas of Hamburg, Berlin, and Munich (see Figure 6a). The image on the right (Figure 6b) shows the DIVI registry hospitals with vulnerabilities on the map. It is easy to recognize that hospitals in both metropolitan areas and rural areas are affected.

GEOLOCATION OF HOSPITALS, NETWORK SERVICES, AND VULNERABILITIES
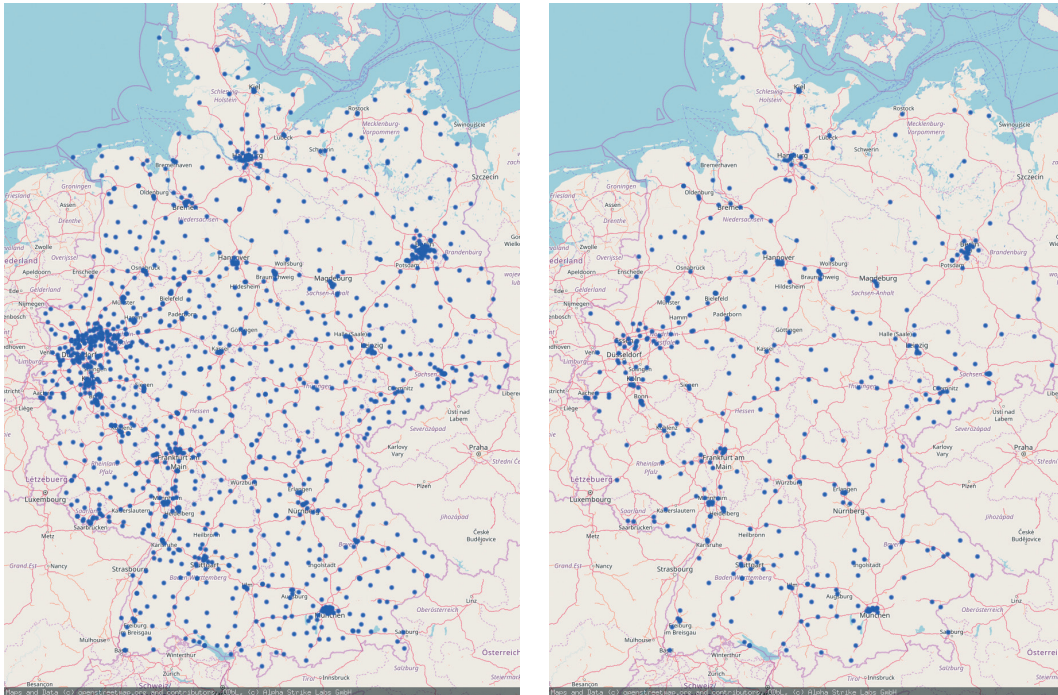
(a) Left side: Identified hospitals and geolocation according to the DIVI registry.
(b) Right side: Identified DIVI hospitals with vulnerabilities.
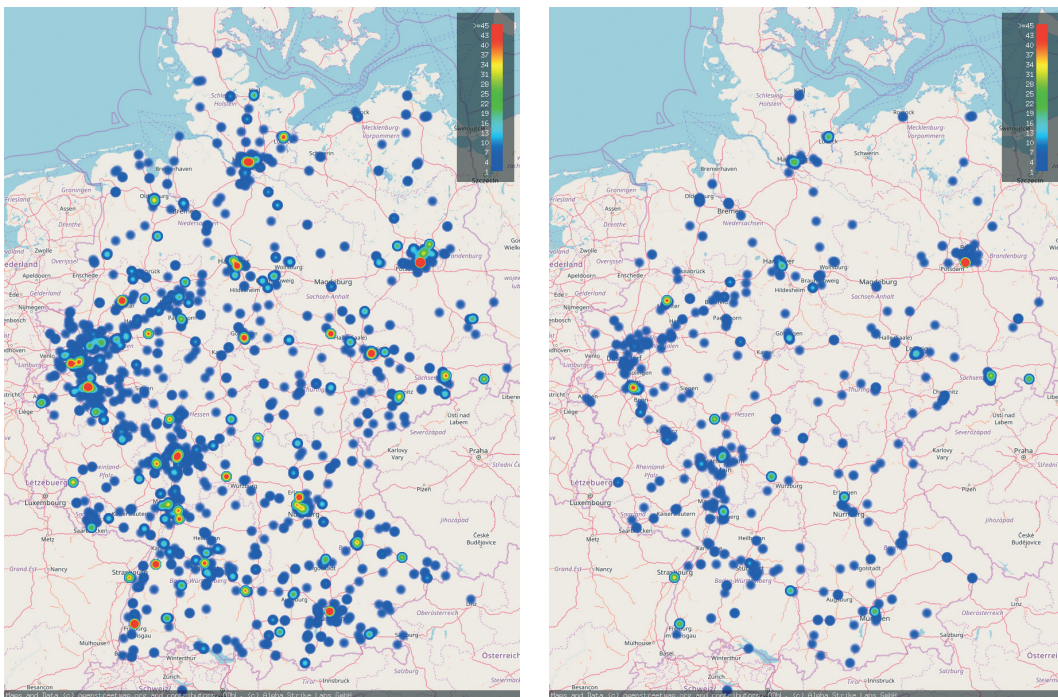
(c) Left side: All network services identified and the approximate Geo-IP location as heatmap.
(d) Right side: Geographical location of the network services with vulnerabilities as heatmap.

In contrast to Figure 6a and 6b, Figure 6c and 6d represent an overview of all 1,555 identified hospitals and their 13,597 network services, which were assigned a geo-coordinate via a Geo-IP resolution. For the Geo-IP resolution, the commercial version of the Maxmind DB [16] with increased resolution was used. Figure 6c shows the network services of all hospitals analogous to Figure 6a, whereas Figure 6d shows only the network services with vulnerabilities.

The main difference between Figure 6a and 6d is that Figure 6a only shows the hospitals of the DIVI Registry and their geographical location. Figure 6d, however, shows a heat map of all identified or vulnerable network services of German hospitals. A comparison of the two graphs clearly reveals that the distribution in the heat map is somewhat smaller, but both graphs show that both rural regions and metropolitan areas have hospitals with vulnerabilities.

First, the following overall CVSS vulnerability statistics should be noted:

TABLE II: CVSS DISTRIBUTION OVERVIEW

CVSS SCORE	Number of vulnerable services
9.0–10 (critical) 7.0–8.9 (high) 4.0–6.9 (medium)	931 443 518
Total vulnerable services:	1,892

Our analysis yielded a total of 1,892 vulnerable services, with nearly half of the vulnerable services carrying a CVSS score of 9 or 10, thereby potentially containing critical vulnerabilities, depending on their version number.

Next, we explore if there is any correlation between the number of identified CVEs (Common Vulnerabilities and Exposures, a reference method for publicly known information-security vulnerabilities and exposures) and the size of the clinical institution. The allocation of the number of beds was taken from the German Hospital Register [8]. An examination of the hospitals with vulnerabilities in relation to their bed capacity shows that hospitals with vulnerabilities represent a total of about 167,000 beds. This represents 32 percent of the approx. 520,000 available hospital beds in Germany. Figure 7 shows the number of identified CVEs in relation to the size of the respective hospitals based on the number of beds. For a better visualization, only hospitals with up to 1,800 beds are drawn; there are only a few facilities with more beds.

Since, naturally, there are more smaller hospitals, there are more data points in the left-hand area of the figure. For better visibility, a detailed representation of this area is shown in Figure 7b.

FIGURE 7: NUMBER OF VULNERABILITIES IN IT SYSTEMS IN HOSPITALS
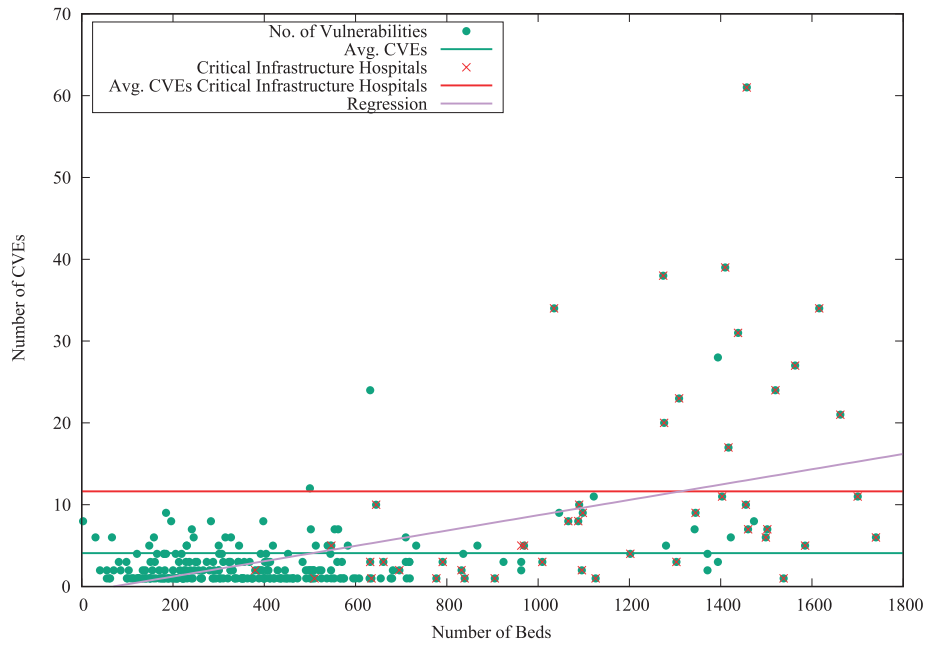
(a) Number of vulnerabilities in hospitals contrasted with the number of beds.
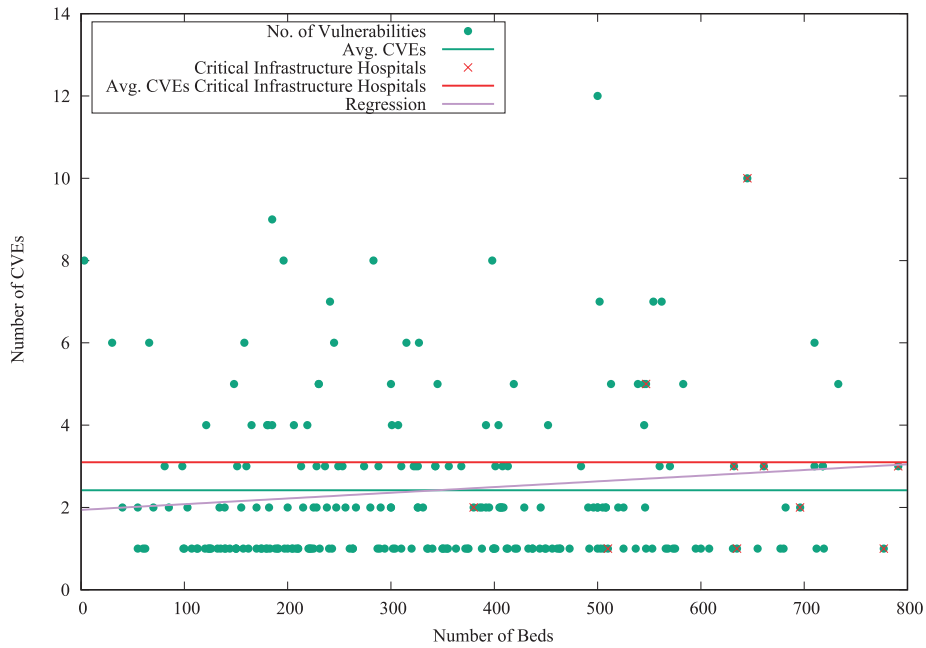
(b) Detail view of the number of vulnerabilities in hospitals with up to 800 beds.

A first look at the data initially reveals an unsurprising trend: As the number of beds increases, so does the number of vulnerabilities found in the IT environments of the respective hospitals. This can probably be explained by the fact that larger hospitals with more beds also typically have more specialized medical departments and corresponding IT equipment, which thus increases the number of IT devices and services and thus the potential attack surface. The regression line of this increase is drawn in the figures correspondingly.

But if we now look at the detailed view in Figure 7b, we notice that a corresponding increase in vulnerabilities in IT systems is much lower among hospitals with up to 800 beds.

Here hospitals in all size ranges have varying numbers of vulnerabilities, without any discernible pattern. We might infer from this that at smaller hospitals, the number of existing vulnerabilities is more likely to depend on the quality of the respective IT service providers, or on specific software products.

With respect to the significantly increasing numbers of vulnerabilities at large hospitals, however, especially at those with more than 1,000 beds, it is apparent that these affect university hospitals in particular. This suggests that the higher CVE figures also reflect the need for more systems and, especially in the research sector, more diversified IT systems and customized or less commonly available software.

With respect to German legislation, the data in Figure 7a offers yet another perspective for analysis: due to the special need for protection of the basic services necessary for modern society, such as electricity water supply, telecommunications, and healthcare, the Federal Office for Information Security (Bundesamt für Sicherheit in der Informationstechnik, BSI) criticalness regulation (KRITIS Act [7]) defines facilities in Germany that are obligated to implement minimum standards and measures in accordance with the BSI Act [2] to ensure sufficient IT security. In the healthcare sector, facilities with more than 30,000 in-patient cases per year are considered critical infrastructure.

Thus an interesting question arose: are higher liabilities resulting from the KRITIS Act reflected in a lower visible CVE attack surface?

In order to evaluate this, the facilities that belong to KRITIS based on the number of cases according to the German Hospital Register [8] were marked accordingly in Figure 7a. Of course, large facilities such as university hospitals fall into this category, but so do some other, smaller facilities. Surprisingly, while the aforementioned

accumulation of vulnerabilities can be seen at university hospitals, smaller institutions also feature systems with more vulnerabilities than average.

For example, an average of 11.63 CVEs was identified for hospitals up to 1,800 beds belonging to KRITIS, while the average value for all of the hospitals up to 1,800 beds analyzed is 4.08. A similar picture emerges when looking at the detail section of smaller hospitals in Figure 7b. While the average number of CVEs at the KRITIS hospitals is 3.1, all analyzed hospitals with up to 800 beds have an average of 2.42 CVEs.

6. DISCUSSION OF RESULTS AND CONCLUSION

In our results section, we first want to acknowledge the known limitations and constraints of our analysis, beginning with the number of vulnerable services our DCS identified (1,892). Firstly, it must be noted that vulnerability identification is done fully automatically through service and banner mapping and CVE entry. In cases where patches have been backported or the administrator has arbitrarily changed banner information, the given CVE match indication naturally would not reflect the actual vulnerability state. For cases like this, we coined the term "Schrödinger vulnerability," which is explained later in this section. Therefore, it may be assumed that the overall number might be a bit lower due to backports or banner changes. On the other hand, other attack vectors such as misconfiguration of services or the use of weak passwords, which are still regularly found today and can represent a high risk for an IT system, are not included in our research. Therefore, taking into account these considerations, our results can also be seen as a *lower bound* for the vulnerabilities of the systems, and their effective exposure can be even higher. Secondly, although DCS uses a number of very well-proven port and service identification methods, in cases where fingerprinting fails, this may create a situation where vulnerability identification is not always accurate.

While we have only analyzed IPv4 addresses in the present work, we are working on the implementation for scanning procedures for the IPv6 protocol. Since its address range is no longer (even approximately) completely scannable due to its sheer size, new and different scanning strategies are required here to reduce and optimize the search space. Efforts in this direction are already underway.

However, regarding the results, it is also important to recognize the great advantage of our method, which is typically unproblematic from a legal point of view due to the evaluation of information provided publicly by Internet-facing services only. By contrast, even if actors such as intelligence services of foreign countries are probably

not bothered by this limitation in various cases, performing vulnerability scans to explicitly search for and thereby trigger vulnerabilities without the permission of the owner of the respective service is problematic for us.

In our research using DCS, we would like to coin the term "Schrödinger vulnerability." In quantum mechanics, "Schrödinger's cat" is a thought experiment conceived by the Austrian-Irish physicist Erwin Schrödinger in 1935 in a discussion with Albert Einstein [26]. In the thought experiment, a hypothetical cat can be considered simultaneously alive and dead because it is associated with a random subatomic event that may or may not occur.

In the thought experiment, a closed box contains a cat and an unstable atomic nucleus, which decays with a certain probability within a certain period of time. The decay would trigger the release of poison gas by means of a Geiger counter and thus kill the cat.

The thought experiment is based on the fact that whenever a system can assume two different states, the coherent superposition of the two states then also constitutes a possible state. It is therefore only an actual observation or measurement being conducted that can distinguish the two original states, as the system may assume either one.

Analogous to this thought experiment, we now consider an IT system with a network service (aka the cat) that has a vulnerability according to its transmitted version in the service banner. With that, however, we cannot know whether the administrator (aka the atomic core) of the IT system has provided the network service with a security patch, because many security updates do not update the communicated software version in the service banner. Consequently, the IT system is in a state of superposition, as it is both vulnerable and non-vulnerable at the same time. Only when the system is subjected to a thorough audit – for example, by analyzing the exact version level or patch level – can we distinguish between the original states, and until then, the system may adopt either a state of vulnerability or one of non-vulnerability.

Consequently, we have to call all vulnerabilities, which in the context of this analysis were mostly identified via the service banner, Schrödinger vulnerabilities or potential vulnerabilities, as they put a system in a state of vulnerability and non-vulnerability simultaneously. From the authors' perspective, all Schrödinger vulnerabilities should therefore, until further audited, be considered a cyber risk.

Further distinctions could be made for future work. For example, evaluations could be created to show the proportion of systems that can no longer be supplied with current

software versions or security updates because the software products have already been marked as end-of-support by the manufacturer.

Let us now summarize the results and findings. First of all, when looking at the attack surface of German clinical providers, the analysis reveals many vulnerabilities and quite high CVSS ratings. Looking at the most noteworthy system occurrences from a security point of view reveals, for instance, two Windows XP operating systems (CVSS 10.0/End of Support since 2015!), open Jitsi VideoChat servers (CVSS 6.11), open unauthenticated squid proxies (CVSS 10.0) allowing proxy misuse, outdated Apache and PHP configurations (CVSS 9.8), direct accessible Intelligent Platform Management Interface (IPMI) login pages, Citrix XenAPP remote access (CVSS 10.0), and direct web links to RDP connections (CVSS 9.8), to give just a few concrete examples.

The main designated and essential function of clinical institutions is healthcare and not IT security. However, the data seem to indicate that there is still a need for better attack surface and vulnerability management, as **approximately 32 percent of the analyzed services were determined to be vulnerable to various degrees, and 36 percent of all hospitals showed vulnerabilities.**

As mentioned, through our analysis we can confirm that healthcare institutions are also affected to a certain extent by the issue of legacy services for which end-of-support was announced years ago and for which security updates are therefore no longer provided, increasing security risk.

Unsurprisingly, larger institutions have more IT systems, potentially leading to a larger attack surface; this was clearly visible in our analysis as well.

Finally, a rather interesting result of our analysis was the fact that hospitals belonging to German critical infrastructure, indicated through their assignment from the KRITIS Act, had **notably higher-than-average vulnerability figures, based on CVE numbers**, among the hospitals we analyzed. We found this result striking, as we had assumed that KRITIS hospitals and clinics would have a much better IT security posture, resulting in lower average CVE numbers than the other hospitals, as they are designated as being critical.

Our analysis concludes that even in 2020, despite its increased criticalness and increased regulation efforts, the German healthcare sector, unfortunately, presented and contained a certain visible amount of attack surface. This attack surface may translate into a national security risk if abused systematically by an intelligent adversary. It is therefore advisable from a state-level risk management perspective to

regularly conduct reconnaissance in cyberspace on all organizations that have been determined critical or essential for a nation.

REFERENCES

[1] Cyber and Infrastructure Agency (CISA), "CISA: Alert (AA20-302A) Ransomware Activity Targeting the Healthcare and Public Health Sector." Accessed: Nov. 15, 2020. [Online]. Available: https://us-cert.cisa.gov/ncas/alerts/aa20-302a

[2] *Act on the Federal Office for Information Technology (BSI Act – BSIG)*. Accessed: Jan. 6, 2021. [Online]. Available: https://www.bsi.bund.de/EN/TheBSI/BSIAct/bsiact_node.html

[3] Jim Bates, "Trojan Horse: AIDS Information Introductory Diskette Version 2.0," *Virus Bulletin*, vol. 6, pp. 1143–1148 (1990).

[4] Coalition Inc., "Cyber Insurance Claims Report H1 2020." Accessed: Feb. 26, 2021. [Online]. Available: https://info.coalitioninc.com/rs/566-KWJ-784/images/DLC-2020-09-Coalition-Cyber-Insurance-Claims-Report-2020.pdf

[5] "Common Vulnerability Scoring System SIG." Accessed: Jan. 6, 2021. [Online]. Available: https://www.first.org/cvss

[6] Andrada Fiscutean, "A history of ransomware: motives and methods behind evolving attacks," *CIO East Africa*, July 28, 2020. Accessed: Jan. 6, 2021. [Online]. Available: https://www.cio.co.ke/a-history-of-ransomware-the-motives-and-methods-behind-these-evolving-attacks

[7] Deutsche Krankenhaus TrustCenter und Informationsverarbeitung GmbH (DKTIG), "BSI-Kritisverordnung (BSI-KritisV)." Accessed: Jan. 4, 2021. [Online]. Available: https://www.gesetzeim-internet.de/bsi-kritisv

[8] Deutsche Krankenhaus TrustCenter und Informationsverarbeitung GmbH (DKTIG), "Deutsches Krankenhaus Verzeichnis." Accessed: Jan. 3, 2021. [Online]. Available: https://www.deutscheskrankenhaus-verzeichnis.de

[9] Robert Koch Institut, "DIVI Intensivregister." Accessed: Jan. 4, 2021. [Online]. Available: https://www.intensivregister.de/#/index

[10] "Cooperative Association for Internet Data Analysis (CAIDA)." Accessed Jan. 3, 2021. [Online]. Available: https://www.caida.org

[11] Jessica Davis, "The 10 biggest healthcare data breaches of 2020," *Health IT Security News*, Dec. 10, 2020. Accessed: Jan. 4, 2021. [Online]. Available: https://healthitsecurity.com/news/the-10-biggest-healthcare-data-breaches-of-2020

[12] Johannes Klick, "Towards Better Internet Citizenship: Reducing the Footprint of Internet-wide Scans by Topology Aware Prefix Selection," Proceedings of the 2016 Internet Measurement Conference, pp. 421–427 (2016).

[13] Forescout Research Labs, "Connected Medical Device Security: A Deep Dive into Healthcare Networks," Tech. Rep. Forescout Technologies, Inc., 2020. Accessed: Dec. 23, 2020. [Online]. Available: https://www.forescout.com/company/resources/connectedmedical-device-security-a-deep-dive-into-healthcare-networks

[14] Joanne Lim, "An Engineering Disaster: Therac-25," 1998. Accessed: Apr. 7, 2021. [Online]. Available: https://www.bowdoin.edu/~allen/courses/cs260/readings/therac.pdf

[15] GLEG Ltd., "Gleg Security @GlegExploitPack." Accessed: Jan. 6, 2021. [Online]. Available: https://twitter.com/GlegExploitPack

[16] Maxmind, "GeoIP Databases and Services." Accessed: Jan. 4, 2021. [Online]. Available: https://www.maxmind.com/en/geoip2-services-and-databases

[17] Melissa Hanson, "Swansea Police Department pays ransom to computer hackers," *Boston Globe*, Nov. 19, 2013. Accessed: Jan. 3, 2021. [Online]. Available: https://www.bostonglobe.com/metro/2013/11/19/swansea-police-pay-ransom-open-files-locked-hackers/7bOdi8i7foNkTmdnokMAkP/story.html

[18] Mitchell Clarke and Tom Hall, "It's not FINished: The Evolving Maturity in Ransomware Operations." Accessed: Jan. 4, 2021. [Online]. Available: https://www.blackhat.com/eu-20/briefings/schedule/index.html#its-not-finished-theevolving-maturity-in-ransomware-operations-21500

[19] "National Vulnerability Database." Accessed: Jan 6, 2021. [Online]. Available: https://nvd.nist.gov

[20] Lily Hay Newman, "Ransomware hits dozens of hospitals in an unprecedented wave," *Wired*, October 29, 2020. Accessed: Nov. 17, 2020. [Online]. Available: https://www.wired.com/story/ransomware-hospitals-ryuk-trickbot/

[21] ZMAP Project. "ZGrab Repository." Accessed: Jan 4, 2021. [Online]. Available: https://github.com/zmap/zgrab2

[22] William Ralston, "The untold story of a cyberattack, a hospital and a dying woman," *Wired*, Nov. 11, 2020. Accessed: Nov. 18, 2020. [Online]. Available: https://www.wired.co.uk/article/ransomware-hospital-death-germany

[23] Kim Zetter, "Why hospitals are the perfect targets for ransomware," *Wired*, March 30, 2016. Accessed: Nov. 16, 2020. [Online]. Available: https://www.wired.com/2016/03/ransomware-why-hospitals-are-the-perfect-targets/

[24] U.S. Department of the Treasury, Office of Foreign Assets Control (OFAC), "Advisory on Potential Sanctions Risks for Facilitating Ransomware Payments." Accessed: Jan. 3, 2021. [Online]. Available: https://home.treasury.gov/system/files/126/ofac_ransomware_advisory_10012020_1.pdf

[25] Phil Goodwin and Andrew Smith, "The State of IT Resilience," White Paper, International Data Corporation (IDC), August 2019. Accessed: Feb. 26, 2021. [Online]. Available: https://www.zerto.com/wp-content/uploads/2019/07/State_of_IT_Resilience_2019.pdf

[26] E. Schrödinger, "Die gegenwärtige Situation in der Quantenmechanik," *Naturwissenschaften*, vol. 23, pp. 807–812 (1935) (in German). Accessed: Mar. 8, 2021. [Online]. Available: https://doi.org/10.1007/BF01491891

The Vulnerability of the Financial System to a Systemic Cyberattack

Bobby Vedral
Managing Partner
MacroEagle Capital
PhD candidate, Modern War Studies Department
Buckingham University, United Kingdom
bobby.vedral@macroeagle.com

Abstract: The financial industry is a prime target of cybercriminal activity, mainly due to the nature of its underlying business ('that's where the money is'[1]), the sector's global interconnectedness, and its high level of digitalization. In response, the private sector has invested vast sums into cybersecurity, and regulators have started to worry about systemic risk. The latter comes in two forms. The first is the risk of a successful cyberattack against a specific financial institution 'spilling over' into the broader financial system, hence unintentionally becoming systemic. The second is the national security concern of a systemic cyberattack launched specifically to disrupt the target's financial ecosystem and therefore the real economy. In both cases, the historic evidence is clear: neither type of event has been recorded thus far. Those who consider warnings of systemic cyberattacks to be little more than threat inflation see that as vindication. This paper takes the opposite view and argues that the probability of a systemic cyberattack is significant enough to warrant a higher degree of cross-disciplinary research and preparedness. To support its main argument, this paper proposes a conceptual framework that focuses on answering two key questions. First, are there sufficient known structural vulnerabilities in the financial ecosystem that could be exploited by a willing adversary? And second, are there plausible scenarios that could see an adversarial nation-state launch such an attack? The answer to both is positive.

Given the lack of data, this analysis is largely qualitative, based on discussions with regulators, chief risk officers, academic experts, and the author's own multi-decade experience as an active participant in the financial market.

Keywords: *finance, resilience, systemic risk, vulnerabilities*

---

[1]   This was the reply of 1930s US bank robber Willie Sutton when asked why he robbed banks. He later co-authored a book titled *Where the Money Was*. See FBI History of Famous Cases & Criminals, https://www.fbi.gov/history/famous-cases [accessed 1 March 2021].

1. INTRODUCTION

The **global financial system** lies at the heart of Western liberal democratic market economies, performing many key intermediary functions, such as deposit-taking, lending, capital markets, investments, and payments. As it is at the forefront of globalization, interconnectedness, and digitalization, its reliance on the confidentiality, integrity, and availability of data and systems is mission critical. It is therefore no surprise that national security experts have long predicted the possibility of a cyberattack on the financial system with systemic consequences, one where states would 'suffer greatly from the instability which would befall world markets should numbers be shifted in bank accounts and data wiped from international financial servers'.[2]

'Systemic cyber risk' therefore means a risk of disruption in the financial system with the potential of serious negative consequences for the real economy. This paper differentiates between two types of systemic cyber risks (see Figure 1). The first is one that starts as an idiosyncratic (company-specific) cyberattack, most probably with criminal intent but not intent to cause system-wide damage, but which inadvertently spills over to the wider financial system. This tends to be the main concern of financial regulators, given that empirical evidence points to cybercrime as the main risk. The second is the 'systemic attack', defined as a nation-state or transnational group acting with the political intent to cause severe financial instability in the target's financial markets and thus harm the real economy as well. This tends to be the main concern of the national security establishment and is the main focus of this essay. In addition, this paper defines 'cyberattack' as an event-risk/shock and not as the long-term undermining of an industry through espionage ('slow burn' or 'death by a thousand cuts').[3]

---

[2]   Jordan Schneider, as quoted in P.W. Singer and Allan Friedman, *Cybersecurity and Cyberwar: What Everyone Needs to Know* (Oxford: Oxford University Press, 2014), 191.

[3]   Jason Healey et al., for example, differentiate between three types of crises: slow burn (long-term undermining), exacerbated crisis (when a financial crisis is already in progress), and initiated crisis (when an adversary uses cyber capabilities to create a financial crisis). See Jason Healey, Patricia Moser, Katheryn Rosen, and Adriana Tache, 'The Future of Financial Stability and Cyber Risk', Brookings Institution, October 2018, https://www.brookings.edu/wp-content/uploads/2018/10/Healey-et-al_Financial-Stability-and-Cyber-Risk.pdf.

FIGURE 1: SYSTEMIC CRISIS BY SPILLOVER VS BY INTENT
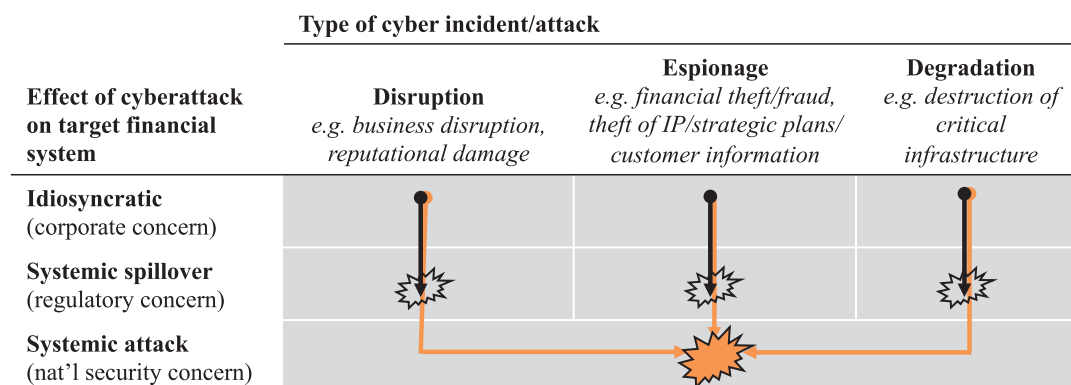
The quantitative evidence regarding systemic cyberattacks is clear: neither a 'systemic spillover' nor a 'systemic attack' have occurred so far. But, as Figure 2 highlights, the financial sector ranks first in most studies when it comes to the frequency of cyber incidents, with most of them idiosyncratic (company specific) and criminal in nature. Also noticeable is that, probably due to the industry's high level of investment in cybersecurity, the average cost per incident is low.[4]

FIGURE 2: CROSS-SECTOR ANALYSIS OF CYBER INCIDENT FREQUENCY AND LOSSES[5]

Category	Frequency of incidents (% of total)	Total loss (% of total)	Mean loss in USD (%ile)	Standard deviation of loss in USD (%ile)
Finance & insurance	24%	16%	USD 1.69 m (10th %ile)	USD 15.45 m (13th %ile)
Most exposed sector	Finance (24%)	Professional, scientific, technical USD 8,778 m (22%)	Transportation and storage USD 16.8 m (100th %ile)	Wholesale trade USD 120.6 m (100th %ile)

This lack of systemic attacks can be attributed to three factors. First, even criminal nation-state actors, such as North Korea, need the capitalist financial system to work in order to cash out. Second, even strategic rivals, like China, need Western capitalist resources to fund their own growth; hence they have no interest in 'biting the hand that feeds them'. And third, systemic attacks on less well guarded critical national infrastructures (CNIs) may be easier to execute.

---

4   An excellent database for cyber incidents in the financial sector is kept by the Carnegie Endowment's 'Timeline of Cyber Incidents Involving Financial Institutions', https://carnegieendowment.org/specialprojects/protectingfinancialstability/timeline [accessed 5 January 2021].

5   For a recent global cross-sector study of cyber incidents in terms of frequencies and losses, see Iñaki Aldasoro, Leonardo Gambacorta, Paolo Giudici, and Thomas Leach, *The Drivers of Cyber Risk*, Bank of International Settlements (BIS), Working Paper No 865, May 2020, https://www.bis.org/publ/work865.htm. All loss data are in millions of US dollars (USD). Twenty sectors and 115,415 incidents are considered.

Why, then, worry about a systemic cyberattack on the financial system? To answer this question, this paper suggests a conceptual framework which defines the probability adjusted economic cost (PAEC) of such an event as a function of the expected economic cost (EEC) should it occur, times the probability of such a systemic cyberattack succeeding, i.e., the probability of a successful attack (PSA). The PSA in turn is a function of: (1) the number of structural vulnerabilities in the financial system that could be exploited; (2) the probability that an adversary has the technical ability to exploit them; (3) the probability that an adversary has the political intent to launch such an attack.

$$PAEC = EEC \; x \; PSA \; (vulnerabilities, ability, intent)$$

Based on various conversations with financial regulators and practitioners, many agree that the key parameter in this model is 'intent'. As Tim Maurer writes, 'the main variable determining whether an actor can cause harm is not technical sophistication, not knowledge of specific vulnerabilities or development of sophisticated codes, but intent. If the intent is there, the capability will follow'.[6] Backed by the above-mentioned absence of precedent for historic systemic attacks, many practitioners point to the lack of intent as the main reason. As a chief information security officer at a major European bank wrote:

> […] the Chinese have zero interest in doing anything destructive to us or any other member of a financial system that makes them wealthy and allows them to wield political and economic influence abroad. Even Iran was circumspect in 2013 when they DDOSed US banks – the attack tech was pretty considerable, but the targets (retail banking websites) were fairly trivial. As long as GDP is a meaningful indicator to a nation-state, I don't believe that nation-state would perpetrate systemic attacks. That said, I'm sure they're curious what their rich citizens are up to, especially if that wealth could be used to aid the opposition, so it wouldn't surprise me if nation-states use espionage tactics against banks. But I can't get my head around any country just wanting to watch the system burn – even North Korea, now that they've discovered how to raise hard currency through hacking.[7]

Hence the focus of this paper is to make the case that the probability of a systemic attack is neither 'zero' nor 'very low', as the historical precedent and consensus view, respectively, imply. The argument is developed in five parts. Section 2 reviews the existing literature on systemic risk in the financial system, which broadly agrees with the assessment that the impact of such an event would be significant and that the

---

6   Tim Maurer, *Cyber Mercenaries: The State, Hackers, and Power* (Cambridge: Cambridge University Press, 2018), 10.
7   Chief Information Security Officer (CISO) of major Western bank, email to author, 22 December 2020.

probability is not zero. Section 3 makes the point that sufficient known vulnerabilities in the current financial ecosystem exist that could be exploited if the will to do so were there. Section 4 addresses the key question about political intent from various perspectives, including historical, cultural, and doctrinal. Section 5 concludes with some basic recommendations and suggestions for further research.

2. LITERATURE REVIEW ON 'SYSTEMIC CYBER RISK' TO THE FINANCIAL SYSTEM

Interest in 'systemic risk' took off after the Great Financial Crisis (GFC) of 2007–2008, although the focus was always more on quantifiable financial aspects, such as market, credit, and liquidity risk. Cyber risk, a sub-category of operational risk, received relatively little attention. With no commonly accepted definition of systemic risk, by **2009** the Financial Stability Board (FSB) outlined three criteria: size, substitutability, and interconnectedness.[8]

By **2013**, and following the Stuxnet disclosures, the White House issued Executive Order 13636, instructing the Department of Homeland Security (DHS) to identify those financial institutions for which a 'cyber incident would have far reaching impact on regional or national economic security'.[9] This led three years later to the creation of the Financial Systemic Analysis & Resilience Centre (FSARC), one of the first collaborative efforts in the private sector.

Judging by the Bank of England's (BOE) semi-annual Systemic Risk Survey (see Figure 3), 'cyberattacks' started to become prominent among financial risk practitioners in **2014**, after the cyberattack on JP Morgan. This attack, widely attributed to Iran, affected over 83 million customers.[10]

---

[8]   Financial Stability Board (FSB), 'Guidance to Assess the Systemic Importance of Financial Institutions, Markets and Instruments: Initial Considerations', IMF-BIS-FSB, October 2009, https://www.fsb.org/wp-content/uploads/r_091107d.pdf.

[9]   US Government, Executive Order No. 13636, 3 C.F.R. 13636 (2013), as mentioned in Jason Healey et al., 'The Future of Financial Stability and Cyber Risk'.

[10]   See, for example, Reuters, 'JP Morgan Hack Exposed Data of 83 Million, Among Biggest Breaches in History', 3 October 2014, https://uk.reuters.com/article/us-jpmorgan-cybersecurity/jpmorgan-hack-exposed-data-of-83-million-among-biggest-breaches-in-history-idUKKCN0HR23T20141003.

FIGURE 3: BOE SYSTEMIC RISK SURVEY – SOURCES OF RISK TO THE UK FINANCIAL SYSTEM[11]
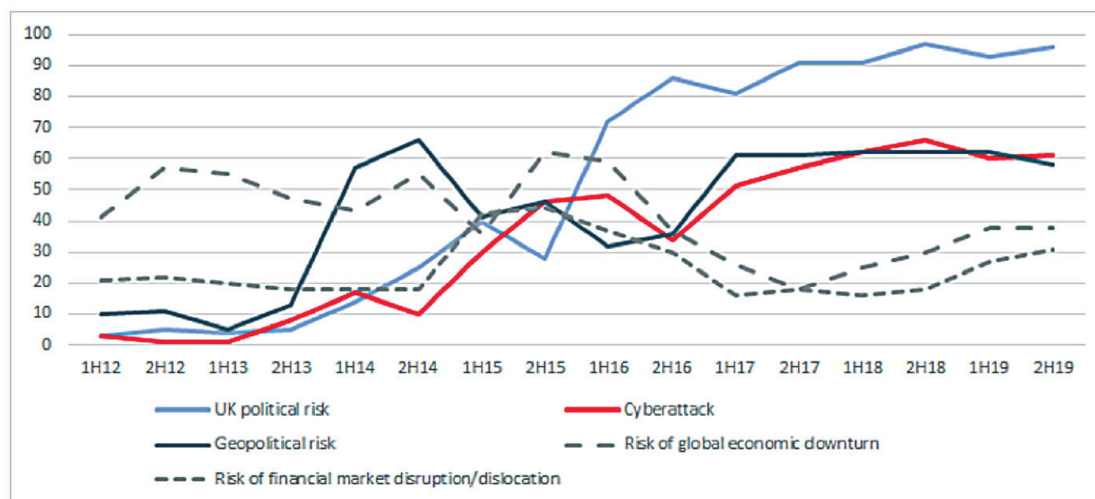
In 2016, the year North Korea attempted to steal USD 951 million from Bangladesh's central bank,[12] the members of the G7 released the *G7's Fundamental Elements of Cybersecurity for the Financial Sector*, suggesting eight elements to follow in designing and implementing a cybersecurity program.[13] Although few academics by that time challenged the view that cyberattacks posed a systemic risk, one important exception was a 2016 *Vox* article by Danielsson et al. The article claimed that systemic cyber crises were extremely unlikely, as most cyberattacks were micro-prudential (company-specific) in nature and required extremely fortunate timing to become systemic.[14]

In 2017, the year of the WannaCry ransomware attack and Equifax hack, the International Monetary Fund (IMF) published a paper describing cyber risk as a textbook example of systemic financial stability risk and identified the main sources of vulnerabilities as access, concentration risk, correlation risk, and contagion risk.[15] Furthermore, the Institute of International Finance (IIF) published a paper that focused on the main types of scenarios that could have systemic repercussions, such as attacks

11   Bank of England (BOE), 'Systemic Risk Survey Results', 2015 H2, https://www.bankofengland.co.uk/systemic-risk-survey/2015/2015-h2; and 2019 H2, https://www.bankofengland.co.uk/systemic-risk-survey/2019/2019-h2. Note: Respondents were asked which five risks they believed would have the greatest impact on the UK financial system if they were to materialize. Answers were provided in free format and subsequently coded into the above categories by the BOE.

12   Jim Finkle, 'Cyber Security Firm: More Evidence North Korea Linked to Bangladesh Heist', Reuters, 3 April 2017, https://www.reuters.com/article/us-cyber-heist-bangladesh-northkorea-idUSKBN1752I4 [accessed 20 December 2020].

13   G7, 'G7 Fundamental Elements of Cybersecurity for the Financial Sector', 11 October 2016, http://www.g7.utoronto.ca/finance/cyber-guidelines-2016.html [accessed 20 December 2020].

14   Jon Danielsson, Morgan Fouche, and Robert Macrae, 'Cyber Risk as Systemic Risk', *Vox*, 10 June 2016, https://voxeu.org/article/cyber-risk-systemic-risk.

15   Emanuel Kopp, Lincoln Kaffenberger, and Christoph Wilson, 'Cyber Risk, Market Failures, and Financial Stability', International Monetary Fund (IMF) Working paper No. 17/185, 7 August 2017, https://www.imf.org/en/Publications/WP/Issues/2017/08/07/Cyber-Risk-Market-Failures-and-Financial-Stability-45104.

on FMI, data corruption, failure of wider infrastructure, and loss of confidence.[16] Finally, the US Office of Financial Research (OFR) identified the three key financial stability risks posed by cyberattacks: lack of substitutability, loss of confidence, and loss of data integrity.[17]

By **2018** the BOE published two important papers. One warned that 'just because there has not been a clear example of a systemic impact on the sector yet, it does not mean it cannot or will not happen in the future'.[18] The second indicated a new and innovative regulatory approach in which the BOE considered the management of operational resilience to be most effectively addressed by focusing on business services rather than on systems and processes. It also announced a new regime of closer cooperation with the security services, as the lack of data required it to rely more on expert judgements.[19]

The same year also saw the publication of a widely cited Brookings paper by Jason Healey et al. identifying the three main differences between cyber and financial shocks (timing, complexity, and adversary intent) and flagging four major concerns: attacker sophistication, single points of failure, international coordination, and new technologies.[20]

Finally, that year the FSB published a 'cyber lexicon' to establish a common language and ensure consistent data collection and reliable measurement.[21] This was followed in **2019** by the International Organization of Securities Commissions (IOSCO) publishing an overview of existing frameworks for cyber regulation to serve as guidance for good practise.[22]

In **2020** the European Systemic Risk Board (ESRB) published two important and related papers, both with substantial input from the BOE. The first paper presents a conceptual model that analyses a cyber incident in four distinct phases: context,

---

16   Martin Boer and Jaime Vazquez, 'Cyber Security and Financial Stability: How Cyber-Attacks Could Materially Impact the Global Financial System', Institute of International Finance (IIF), September 2017, https://www.iif.com/Portals/0/Files/IIF%20Cyber%20Financial%20Stability%20Paper%20Final%20 09%2007%202017.pdf?ver%3D2019-02-19-150125-767.

17   Office of Financial Research (OFR), 'Cybersecurity and Financial Stability: Risks and Resilience', OFR Viewpoint 17-01, 15 February 2017, https://www.financialresearch.gov/viewpoint-papers/files/ OFRvp_17-01_Cybersecurity.pdf.

18   Phil Warren, Kim Kaivanto, and Dan Prince, 'Could a Cyber-Attack Cause a Systemic Impact in the Financial Sector?' Bank of England (BOE), *Quarterly Bulletin*, Q4 2018, https://www.bankofengland. co.uk/-/media/boe/files/quarterly-bulletin/2018/could-a-cyber-attack-cause-a-systemic-impact-final-web.pd f?la=en&hash=61555F2E3C15AD6B65E845C13238733B9364D4F6.

19   Bank of England (BOE), 'Building the UK Financial Sector's Operational Resilience', Discussion Paper, BOE-PRA-FCA, July 2018, https://www.bankofengland.co.uk/-/media/boe/files/prudential-regulation/ discussion-paper/2018/dp118.pdf.

20   Healey et al., 'The Future of Financial Stability and Cyber Risk'.

21   Financial Stability Board (FSB), 'Cyber Lexicon', 12 November 2018, https://www.fsb.org/2018/11/cyber-lexicon/.

22   International Organization of Securities Commissions (IOSCO), 'Cyber Task Force – Final Report', June 2019, https://www.iosco.org/library/pubdocs/pdf/IOSCOPD633.pdf.

shock, amplification, and systemic event. It then uses the model and discusses three hypothetical scenarios: (1) the incapacitation of a large domestic bank's payment system; (2) the malicious destruction of account balance data; (3) the scrambling of price and position data.[23] In the second paper, the same model is reviewed and an extensive number of systemic mitigants are listed.[24] In December, the Carnegie Endowment published a report on systemic cyber risk, identifying and providing detailed recommendations for six priority areas: cyber resilience, international norms, collective response, workforce challenges, capacity-building, and digital transformation.[25]

In summary, the existing literature shows that systemic cyber risk is a concern for financial regulators, especially those in Britain and the US, where most of the relevant publications originate from. It is also noticeable that the concern is fairly recent; most of the more in-depth studies have been produced over the last one or two years. The current paper aims to build on the existing literature in that it focuses specifically on the likelihood of a systemic attack launched by an adversarial nation-state with the intent to disrupt the target financial system. To address this question, this paper will now turn towards highlighting a number of structural vulnerabilities in the global financial system that could be exploited as either a target or an amplifier during such an attack. This goes back to this paper's conceptual model: that the probability of success is conditioned in part on the availability of vulnerabilities to exploit.

3. STRUCTURAL VULNERABILITIES IN THE FINANCIAL ECOSYSTEM

This section provides an overview of 10 known structural vulnerabilities of the financial ecosystem that highlight liberal democracies' higher exposure to financial instability due to differences in their respective political economies (openness, values), structural concentration risks (currency, geography, counterparty, participants, strategy) or amplification channels (technology, trust) across the system. The list is not meant to be exhaustive or an in-depth analysis of any one vulnerability. The intention is to highlight the fact that there is no shortage of them and that the number of possible vulnerabilities is, if anything, a parameter that increases the PSA factor in the conceptual model.

23 European Systemic Risk Board (ESRB), 'Systemic Cyber Risk', February 2020, https://www.esrb.europa. eu/pub/pdf/reports/esrb.report200219_systemiccyberrisk~101a09685e.en.pdf.

24 Greg Ros et al., 'The Making of a Cyber Crash: A Conceptual Model for Systemic Risk in the Financial Sector', European Systemic Risk Board (ESRB), Occasional Paper Series, No 16, May 2020, https://www. esrb.europa.eu/pub/pdf/occasional/esrb.op16~f80ad1d83a.en.pdf.

25 Tim Maurer and Arthur Nelson, 'International Strategy to Better Protect the Financial System against Cyber Threats', Carnegie Endowment for International Peace, 2020, https://carnegieendowment.org/files/ Maurer_Nelson_FinCyber_final1.pdf.

1 – Degree of financial openness. Figure 4 compares four autocratic regimes with the main Western financial centres (US, UK) and ranks them based on military and socioeconomic criteria. Although autocratic states differ greatly in terms of economic size, they show a much tighter control over their media and financial systems, which suggests a greater degree of control in times of crisis. For example, although China has the four largest banks by assets in the world, their international expansion is minimal.[26] This contrasts with their American and European peers, who have extensive international networks. Or take North Korea, which has a record of attempting to paralyse financial networks in South Korea through cyberattacks, but whose own financial system is largely analogue and hence immune.[27]

FIGURE 4: KNOW YOUR ADVERSARY (COUNTRY'S GLOBAL RANKING BY CATEGORY)

Country	Cyber Power[28] (2020)	GDP[29] (2019)	Military Spending (2019)[30]	Press Freedom[31] (2020)	Financial Openness[32] (2018)
US	1	1	1	45	1
UK	3	6	8	35	1
China	2	2	2	177	105
Russia	4	11	4	149	85
Iran	23	29	18	173	165
North Korea	16	no data	no data	180	no data

2 – Domestic politics. Given the international exposure of Western financial institutions, it is likely that they are more vulnerable to political pressure generated by domestic conflicts, such as when consumer activism at home clashes with commercial interests overseas. For example, Beijing's 2020 imposition of a new security law in Hong Kong saw the British government lead the international condemnation, while HSBC and Standard Chartered, two British banks with significant commercial

26   Ali Zarmina, 'The World's Largest 100 Banks, 2020', S&P Global Market Intelligence, 7 April 2020, https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/the-world-s-100-largest-banks-2020-57854079.

27   As mentioned in Kong Ji Young, Lim Jong In, and Kim Kyoung Gon, 'The All-Purpose Sword: North Korea's Cyber Operations and Strategies', *11th International Conference on Cyber Conflict: Silent Battle* (Tallinn: NATO CCDCOE, 2019), 151.

28   Belfer Center, 'National Cyber Power Index 2020', Harvard Kennedy School, September 2020, https://www.belfercenter.org/sites/default/files/2020-09/NCPI_2020.pdf.

29   GDP data from 'World Development Indicators' databank, World Bank, https://databank.worldbank.org/source/world-development-indicators [accessed 30 December 2020].

30   Stockholm International Peace Research Institute (SIPRI), 'Trends in World Military Expenditure', April 2020, https://www.sipri.org/sites/default/files/2020-04/fs_2020_04_milex_0_0.pdf. Military spending measured in billions of US dollars.

31   Reporters Without Borders (RSF), '2020 World Press Freedom Index' dataset, https://rsf.org/en/ranking [accessed 20 December 2020].

32   The Chinn-Ito Financial Openness Index (KAOPEN) is an index measuring a country's degree of capital account openness and has been updated to 2018. The reference paper is Menzie D. Chinn and Hiro Ito, 'What Matters for Financial Development? Capital Controls, Institutions, and Interactions', *Journal of Development Economics* 81, no. 1 (October 2006): 163–192. The dataset is available under http://web.pdx.edu/~ito/Readme_kaopen2018.pdf.

interests in China, publicly endorsed the new law.[33] The point here is not to judge if Western institutions should have these conflicts but to highlight that they exist and to encourage further research into their implications.

3 – Currency concentration. Figure 5 provides a snapshot of the currency market, where USD 6.6 trillion is traded every day.[34] The US dollar is strongly overrepresented (when compared to US GDP), while the Chinese yuan is strongly underrepresented (when compared to China's GDP). While in the short term, this may seem to confer an advantage on the US – for instance, to be able to apply economic sanctions on countries such as Russia and Iran – there are three drawbacks. First, any loss of confidence in the US dollar would immediately have systemic repercussions. Second, the sanctions have driven Russia and China to develop their own parallel financial infrastructure, which will increase their operational independence and resilience in the future.[35] Third, a country falling under US dollar sanctions is so cut off from the global financial system that it might consider there to be no downside in attacking the system.

FIGURE 5: US DOLLAR HEGEMONY IN THE FINANCIAL SYSTEM

	% GDP (2019)[36]	Daily currency turnover, % of total (2019)	Currency as % of global reserves[37]
United States (USD)	24.4%	44.1%	60.4%
China[38] (RMB)	16.3%	2.1%	2.1%
Euro Area (EUR)	15.2%	16.1%	20.5%
All others	54.9%	37.7%	17.0%

4 – Geographic concentration. The global financial system is extremely concentrated in two markets: the US (New York), mainly for capital raising, and the UK (London), mainly for international banking, such as currency and derivative transactions. While this has clear advantages such as the clustering of expertise, it also has a major drawback

---

33   BBC, 'HSBC and StanChart Back China Security Laws for HK', 4 June 2020, https://www.bbc.co.uk/news/business-52916119.

34   Bank for International Settlements (BIS), 'Foreign Exchange Turnover in April 2019', Triennial Central Bank Survey, 16 September 2019, https://www.bis.org/statistics/rpfx19_fx.pdf.

35   See, for example, Russia Briefing, 'Russian and Chinese Alternatives for SWIFT Global Banking Network Coming Online', 17 June 2019, https://www.russia-briefing.com/news/russian-chinese-alternatives-swift-global-banking-network-coming-online.html/.

36   'GDP (Current USD)', as per World Development Indicators, World Bank, https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?year_high_desc=true [accessed 2 July 2020].

37   'Currency Composition of Official Foreign Exchange Reserve - At a Glance - IMF Data', currency composition as per Q3 2020, IMF Currency Composition of Official Foreign Exchange Reserves (COFER) database, https://data.imf.org/?sk=E6A5F467-C14B-4AA8-9F6D-5A09EC4E62A4 [accessed 20 December 2020].

38   These numbers exclude Hong Kong SAR and the Hong Kong dollar (HKD).

in that it offers obvious geographic targets. It is yet to be seen if the pandemic-induced trend toward remote working will endure and help reduce this vulnerability.

FIGURE 6: GEOGRAPHICAL DISTRIBUTION OF TOP FIVE FOREIGN EXCHANGE AND INTEREST RATE DERIVATIVES TURNOVER

Country	Equities[39]	FX turnover[40]	IR Derivatives[41]
United States	54.5%	26.5%	32.2%
Japan	7.7%	4.5%	1.7%
United Kingdom	5.1%	43.1%	50.2%
China (incl. Hong Kong)	4.0%	8.2%	6.0%
France	3.2%	2.0%	1.6%

5 – Central counterparty concentration. One of the key objectives of the regulatory reform efforts after the Great Financial Crisis (GFC) of 2007–2008 was to move from a trading ecosystem centred on banks and bespoke bilateral contracts to one where exchanges, central counterparties (CCPs), and standardized contracts take centre stage (see Figure 7). But while connecting firms through centralized networks makes sense, when market and liquidity risk are a regulator's key priority, it might have inadvertently created a single point of failure from an operational perspective.

FIGURE 7: SECURITIES TRADING ECOSYSTEM BEFORE AND AFTER THE GREAT FINANCIAL CRISIS (GFC)
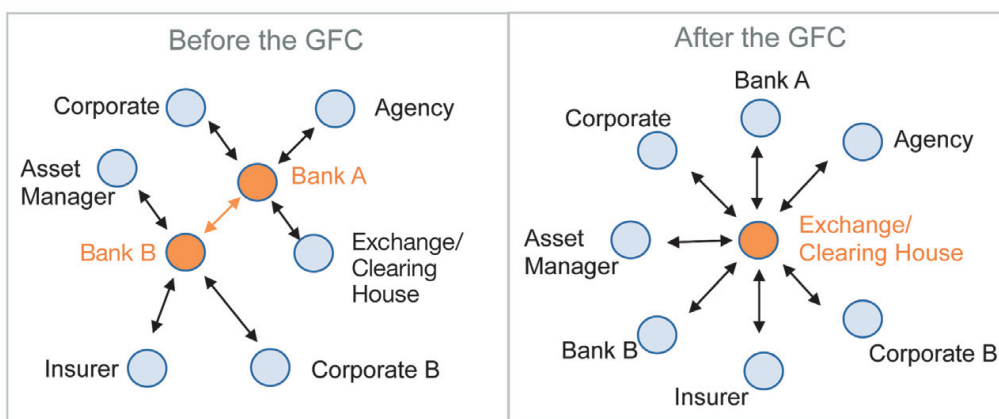
39 Statista, 'Distribution of Countries with Largest Stock Markets Worldwide by Share of Total World Equity Market Value', January 2020, https://www.statista.com/statistics/710680/global-stock-markets-by-country/ [accessed 20 December 2020].
40 BIS, 'Foreign Exchange Turnover'.
41 Bank for International Settlements (BIS), 'OTC Interest Rate Derivatives Turnover in April 2019', Triennial Central Bank Survey, 16 September 2019, https://www.bis.org/statistics/rpfx19_ir.pdf.

6 – Market participant concentration. The financial industry is no exception to the global trend of industry concentration, usually a regulatory concern for reasons of competition and antitrust.[42] Like geographic concentration, this has the advantage of clustering expertise and ability to invest in cybersecurity. But it also means that once broken, the risk of systemic contagion is higher. Also worth considering are the network externalities of smaller financial institutions, which are probably less protected and hence more exposed. A recent Federal Reserve paper showed that under the right circumstances, a single coordinated attack on an average of 24 small institutions could lead to at least one of the top five institutions' reserves dropping below its minimum liquidity.[43]

7 – Investment strategy concentration. In the aftermath of the GFC, as banks and insurance companies de-risked, the asset management industry picked up much of the slack. At the same time, with financial conditions extremely loose (low interest rates and central bank balance sheet expansion), equity markets rose and investors shifted towards passively managed funds, increasing the amount of 'herding', as these funds merely track indices and benchmarks. A recent study by the US Federal Reserve Board noted that this active-to-passive shift meant an increased risk of amplifying market volatility (due to herding) and led to increasing industry concentration (economies of scale).[44] A cyberattack on the integrity of critical market data underlying these benchmark indices and strategies would likely paralyse much of the investment market.

8 – FinTech and Digitalization. FinTech is a relatively new term that, loosely defined, refers to technological innovations that affect financial services. These include cloud computing, robotics, artificial intelligence (AI) and machine learning (ML), mobile applications, big data analytics, blockchain or distributed ledger technology (DLT), cryptography, and quantum computing. While FinTech clearly has the potential to enhance, transform, and disrupt financial services, it also poses significant new risks. First is the risk that speed and innovation comes at the expense of safety. Second is the lack of visibility for regulators to assess technological commonalities.[45] Third is

---

42 See, for example, *Economist*, 'Capitalism is Becoming Less Competitive', 10 October 2018, https://www.economist.com/open-future/2018/10/10/capitalism-is-becoming-less-competitive [accessed 10 December 2020].

43 Thomas Eisenbach, Anna Kovner, and Michael Junho Lee, 'Cyber Risk and the US Financial System: A Pre-Mortem Analysis', Federal Reserve of New York, Staff Reports No 909, January 2020, https://www.newyorkfed.org/research/staff_reports/sr909.

44 Kenechukwu Andau et al., 'The Shift from Active to Passive Investing: Potential Risks to Financial Stability', Federal Reserve Board, Washington, 15 May 2020, https://www.federalreserve.gov/econres/feds/files/2018060r1pap.pdf.

45 Claudia Buch, 'Digitalization, Competition, and Financial Stability', Deutsche Bundesbank, 17 August 2019, https://www.bundesbank.de/en/press/speeches/digitalization-competition-and-financial-stability-799792.

the risk that the rapid adoption of new technology makes existing regulatory models obsolete and hence creates new risks to financial stability.[46]

9 – Automation. In July 2015 the New York Stock Exchange (NYSE) halted the trading of USD 28 trillion worth of stocks because of a coding error at Knight Capital Group, which itself declared bankruptcy a few days later. While this and various other technical flash crashes do not themselves point to anything bigger, they do reveal the fragility of the underlying reliance on high-frequency data-driven systems, quantitative algorithms, and ever-increasing trading speed, which taken together can lead to errors spreading faster and further, outpacing a management's ability to take corrective action. As Lucas Kello points out: 'A major and international interruption of stock-trading platforms could create psychological reverberations that undermine public confidence in the entire financial system.'[47]

10 – Trust. The lifeblood of financial markets is news, data, and trust. Since cyber operations allow attackers to target the integrity and/or availability of key financial data (as mentioned above) or spread misinformation, a cyberattack becomes the weapon of choice, should finance be the target. An early example of the impact of misinformation was the Syrian Electronic Army's takeover of the Associated Press's Twitter account in April 2013, sending the fake message of a bomb attack on President Obama, that caused the Dow to plunge 146 points in a few seconds, erasing USD 136 billion in market value.

As mentioned above, the point of illustrating these vulnerabilities is to flag that the financial system has various vulnerabilities that can be exploited, if the will to do so exists. In the next section, we turn to the crucial question of intent.

4. ON POLITICAL INTENT

As mentioned in the introduction, one of the most consistent pushbacks on the PSA is that most practitioners consider such an act economically irrational and hence conclude that there is little or no chance of an adversary acting this way. Six arguments can be made to argue that the probability is high enough to make the PSA and therefore the PAEC significant.

First, **historical precedent** shows the fallacy of the economic interdependence argument.[48] Henry Kissinger recently warned that the current Sino-American state of

---

[46] Speech by Loretta J. Mester at the 2019 Financial Stability Conference in Cleveland, Ohio, 21 November 2019, https://www.clevelandfed.org/en/newsroom-and-events/speeches/sp-20191121-cybersecurity-and-financial-stability.aspx.

[47] Lucas Kello, *The Virtual Weapon and International Order* (New Haven: Yale University Press, 2017), 124.

[48] Exploring this argument is beyond the scope of this paper, but the roots of the interdependence argument can be found in the early 1970s. See, for example, R.O. Keohane and J.S. Nye, *Power and Interdependence* (Boston: Little, Brown, 1977).

relations bears similarities to the conditions that led to World War I.[49] Back then, well-regarded authors such as J.G. Bloch (*Is War Now Impossible?*) and Norman Angell (*The Great Illusion*) argued that economic interdependence, especially the cross-border flow of credit, technological innovation, and pure self-interest, would triumph in the face of narrow concepts of national interest and hence make war impossible.[50] It did not.

Second, nations with different histories, cultures, geographies, economies, and real or perceived threat perceptions still struggle to correctly assess other nations' **strategic interests**. Recent evidence of this includes the Iraqi invasion of Kuwait (1990), the 9/11 attacks (2001), the ISIS offensive (2014), the Russian invasion of Crimea (2014), the Chinese militarization of the South China Sea (2016), and the recent crackdown in Hong Kong (2020), most of which caught Western intelligence services by surprise. This is relevant, as some Western observers believe that China will not overreact when it comes to Taiwan. But as Coker correctly points out: 'The US palpably failed [...] in its own overreaction to 9/11. There is no "reason" to suspect the Chinese of being any more sophisticated in reasoning out what is in their best interests.'[51]

Third, a common misconception is to see a systemic attack on the financial system as an opening shot to war. However, it could just be an act of **non-violent political coercion** intended to strategically undermine another nation's will to fight by highlighting the economic cost of intervention. To return to the Taiwan example, if China wanted to send a strong message, a cyberattack would probably be preferable to a kinetic attack. As Adam Segal points out: 'In the future the moral expectation may be that states use cyber weapons before kinetic ones.'[52]

Fourth, **military doctrine** naturally evolves with technological capabilities. The 2010 military doctrine of the Russian Federation made clear that information warfare is an instrument 'to achieve political objectives without the utilization of military force'.[53] In a similar fashion, Chinese strategists speak of strategic cyber warfare being intended to 'paralyze state apparatus and [bring] about social unrest and the downfall of enemy countries' governments'.[54] According to Coker:

49   Peter Martin, 'Kissinger Warns Biden of US-China Catastrophe on Scale of WWI', Bloomberg News, 16 November 2020, https://www.bloomberg.com/news/articles/2020-11-16/kissinger-warns-biden-of-u-s-china-catastrophe-on-scale-of-wwi.

50   Lawrence Freedman, *The Future of War: A History* (Great Britain: Penguin, 2018), 42–43.

51   Christopher Coker, *The Improbable War: China, the United States and The Logic of Great Power Conflict* (London: Hurst & Company, 2015), 33.

52   Adam Segal, *The Hacked World Order* (New York: PublicAffairs, 2017), 270.

53   Segal, *The Hacked World Order*, 70.

54   Teng Jianqun and Xu Longdi, *Cyber War Preparedness, Cyberspace Arms Control and the United States* (Beijing: China Institute of International Studies, 2014), 48.

> The use of cyber-attacks is entirely consistent with Chinese strategic thinking. 'Force' ('Li') only appears nine times in Art of War's 13 chapters. As far as Sun Tzi was concerned victory and defeat are essentially psychological. The object is to inflict pain psychologically rather than physically – to put the enemy on the back foot and keep him there.[55]

Fifth, targeting the financial system allows attackers to disproportionally **target the elites**. For example, in the US, the top 10% of households owned 88.1% of stock wealth in the fourth quarter of 2019, the highest level since record-keeping began in 1989.[56] The implication of this is twofold in the case of a coercive cyberattack: either the elites will put pressure on their national government to safeguard their financial interests, or the 'bottom 90%' will put pressure to stop the financial chaos before it spreads into the real economy.

Sixth is a question of **reciprocity**. The US and UK are reported to have 'war-gamed a massive cyber strike to black out Moscow if Vladimir Putin launches a military attack on the West'.[57] One can only assume that, in the unlikely case they had not thought about it already, they have now taken notice and are planning their own measures.

5. CONCLUSION

This paper has argued that the probability of a successful systemic cyberattack (PSA) is higher than the one implied by precedent (zero) or the very low estimate given by various financial practitioners. Given that the economic impact of such an attack (EEC) would most likely be significant, any non-zero PSA implies a high enough probability adjusted economic cost (PAEC) to warrant investment into further research and preparedness planning. In fact, it is possible that the numerous observed cyberattacks on the financial sector are serving as an ongoing laboratory where malicious payloads and exploits are developed and refined in order to be used later for systemic cyberattack purposes.

Future research could consider a number of other questions. For instance, it could attempt to quantify the parameters identified in the conceptual model, where, for example, the EEC should vary from country to country given differences in the underlying economic size and structure. Moreover, an in-depth analysis could be made into any of the mentioned vulnerabilities, not only in terms of their stand-alone

---

55   Coker, *The Improbable War*, 160.
56   Federal Reserve, 'DFA: Distribution Financial Accounts' database, https://www.federalreserve.gov/releases/z1/dataviz/dfa/distribute/chart/ [accessed 20 December 2020].
57   Caroline Wheeler, Tim Shipman, and Mark Hookham, 'UK War-Games Cyber Attack on Moscow', *Sunday Times*, 7 October 2018, https://www.thetimes.co.uk/article/uk-war-games-cyber-attack-on-moscow-dgxz8ppv0.

impact but also considering the potential multiplier effect if two or more were targeted at the same time.

As for basic policy recommendations, three stand out. First, from the publicly available literature review, it is clear that **US and UK financial regulators** are at the forefront in terms of quantitative and qualitative analysis. That makes intuitive sense, since both host the world's major financial centres but also benefit from world-leading cybersecurity and intelligence services. NATO members' financial regulators should actively seek their advice and look for possibilities for cooperation.

Second, the ultimate backup plan against a systemic cyberattack is **to switch off** the digitalized part of the financial system while keeping the real economy running. One European financial regulator feared that the financial industry was too digitalized for this alternative to be an option.[58] But on the other hand, as recently as February 2018, Sweden's central bank governor called for public control over the country's (largely private) payment system, fearing that a fully digital system would be vulnerable to attack. He said: 'It should be obvious that Sweden's preparedness would be weakened if, in a serious crisis or war, we had not decided in advance how households and companies would pay for fuel, supplies and other necessities.'[59] Regulators should therefore consider public backup institutions on zero-trust architecture that, in an act of ultimate resilience, would allow for commercial banking to 'go manual'. A possible analogy is the response of Norsk Hydro to a March 2019 cyberattack: the Norwegian firm averted a major operational disaster by switching its plants to manual.[60] One idea would be to use the military's logistical capabilities to support the financial regulators and the private sector in providing an emergency backup banking system to the real economy during a state of emergency.

Third, cross-disciplinary scenario planning and war-gaming involving practitioners from finance, intelligence services, technology providers, and the armed forces should be encouraged. A common language should be created, and industry-specific jargon should be avoided so as not to create distance and separation in cross-disciplinary communication. Critical issues are too often misunderstood and hence remain undebated. Worst-case-scenario planning between finance, financial regulators, and national security needs to be encouraged, as economic interconnectedness and rational-choice theory are no protection against geopolitical conflict.

---

58   Discussion between the author and a senior European banking representative in charge of operational risk, December 2020.

59   David Crouch, 'Being Cash-Free Puts Us at Risk of Attack: Swedes Turn against Cashlessness', *Guardian*, 3 April 2018, https://www.theguardian.com/world/2018/apr/03/being-cash-free-puts-us-at-risk-of-attack-swedes-turn-against-cashlessness.

60   *Engineer*, 'Norsk Hydro Switches Plants to Manual after Cyber-Attack', 20 March 2019, https://www.theengineer.co.uk/norsk-hydro-cyber-attack/.

Strategic Cyber Effects in Complex Systems: Understanding the US Air Transportation Sector

Charles Harry

Associate Research Professor

School of Public Policy

University of Maryland

College Park, MD, United States

charry@umd.edu

Skanda Vivek

Assistant Professor

School of Science and Technology

Georgia Gwinnett College

Lawrenceville, GA, United States

skanda.vivek@gmail.com

Abstract: US policy-makers have coalesced around the need to develop a risk-based approach for managing strategic effects of cyber attacks. This paper uses graph networks of US air infrastructures from the Department of Transportation to develop the Strategic Disruption Index (SDI), a means to assess the loss of effective transportation network capacity of passengers resulting from various cyber attack scenarios. Dynamic effects are measured using an agent based model to assess the ensuing propagating air passenger delays. Results from this analysis show strategic effects are influenced by airport and airline network structure and induce dynamic effects across the entire sector. We find that the largest national strategic effects are generated through the disruption of key vendor relationships that can potentially affect multiple private operators simultaneously. Policy-makers who are charged with developing means of measuring national risk can apply this approach to evaluate strategic impacts to any number of domestic or international transportation networks. They can also use the approach to compare impacts between disparate infrastructure networks to prioritize resources that best limit the range of strategic risk.

Keywords: *strategic cyber risk, critical infrastructure, complex systems analysis*

1. INTRODUCTION

Governments have increasingly focused on the range of strategic impacts cyber attacks can generate, including significant disruptions to critical infrastructure. US policy-makers have sought to adopt risk-based approaches to cybersecurity to promote resilience in critical infrastructure but have struggled with ways to quantify risk. The Department of Homeland Security notes, "We lack integrated and scalable adoption and application of systemic risk assessment, resulting in ineffective and uncoordinated application of resources for cybersecurity" [1]. This challenge arises in part from the inability to assess strategic impact across many independent but related organizations that support critical public services [2]. The Cyber and Infrastructure Security Agency (CISA) has defined several National Critical Functions (NCF) – key strategic services where a cyber attack could generate a significant public concern. These include the movement of air passengers. We address two key questions in this paper. First, how do you quantify the strategic effects of a cyber attack on airports, airlines, or key vendors that disrupt portions of the passenger air network? Second, which events generate the greatest concern for national operators and policy-makers? To answer these questions, we explore effective capacity loss in passenger transport networks and the resulting propagation delay for connecting flights in US air infrastructure as a measure of strategic impact. The approach discussed in this paper contributes to the literature on strategic effects of cyber attack to air transport systems specifically and provides a means of calculating strategic impact on critical infrastructure more generally.

Disruptions at Delta in 2016, United Airlines in 2017, and Southwest in 2019 have highlighted the growing impact IT failure can have across air networks [3]. While most disruptive events against air infrastructure are a result of unintended consequences of ill-timed application rollouts, a small number of incidents have resulted from malicious actors who have succeeded in disrupting air operations through attacks on proprietary systems at airlines or by attacking airport infrastructure directly. Attacks on RavnAir in 2019 [4], Polish national carrier LOT in 2015 [5], and Russian attacks on Swedish [6] and Ukrainian airport infrastructure highlight the potential for malicious disruptive activity [7]. In this paper, we assume that cyber attacks disrupt systems that are key to transportation of passengers by aircraft. Examples of these types of attacks include disabling of tower communications systems, air passenger booking software, or aircraft weighing systems. The specific technical details of the cyber attack, such as the deployment of a ransomware variant, a specific exploit used, or different persistence mechanisms, is not dealt with specifically, as policy-makers are more interested in the operational effects than in the technical details of the malware. However, some techniques are more likely to occur than others, and their probability of occurrence can be paired with the results from this analysis to generate a measure of cyber risk.

Data from the 2019 US Department of Transportation air carrier statistics are used to construct national graph models of air transportation and regional graphs aligned with the Federal Emergency Management Agency (FEMA)'s response zones. While cyber-induced disruption to US air infrastructure would no doubt create impacts to international flights, we do not explicitly address the effects in this analysis. However, the approach discussed could be applied to international air passenger networks. To assess the loss of effective transport network capacity, we introduce the Strategic Disruption Index (SDI) to measure the weighted capacity loss the disruption would represent on air networks. Finally, we assess the dynamic propagation delay of aircraft and the impacts they would have on connecting flights. Our key findings suggest that the largest strategic effects on the national air network would result from attacks on airline infrastructure and are most concerning when involving the disruption of common-use third-party vendors. Such attacks would substantially reduce air transport capacity at several airports simultaneously, generating substantial delays across the graph structure. This finding suggests that the cyber security of key vendors or operators in air transportation remains a greater strategic vulnerability to local and national air infrastructure. It also suggests, though, that this relationship likely exists in other infrastructures as well.

2. LITERATURE ON CYBER EFFECTS, ATTACKS ON NETWORK STRUCTURE, AND PROPAGATION DELAY

Scholars have attempted to categorize and measure the range of cyber impacts, including estimates of the categories of harm [8, 9] and organizational impact [10], but generally have not linked together the primary (technical), secondary (organization), and second-order (society) effects that bind the actions of a threat actor on a specific device to the cascading impacts on society [11, 12]. Several studies have looked at impacts to critical infrastructures, including the electrical grid [13], water distribution [14], and even transportation [15], yet tend to narrowly focus on defining specific technical vulnerabilities tied to the provision of the service, not on quantifying the capacity loss or delay to provisioning of the service on society. Dieye et al. [16] and Santos et al. [17] come closest in their analysis of macroeconomic linkages but focus their approach on output loss and price changes as a result of the disruption to ports. Their analysis does not measure changes in the capacity of the transport network holistically or assess the delay propagation stemming from attacks on the organizational network infrastructure of the entire sector.

The estimation and description of network structures have been broadly explored in critical transport infrastructures [18]. Exploration of airline routes [19], roads [19, 20], railways [21], and river networks [21, 22] detail the structure of linkage but do

not themselves explore disruptive impacts as a result of cyber attacks on the network structure specifically. Amaral et al. [19] explore the structure of large national airport networks, highlighting their scale-free structure, but do not seek to quantify the impact to flight operations from either natural or man-made disruptions. Other efforts to estimate the impact of vertex removal, specifically in internet structures, note the resilience of scale-free network structures to single vertex removal [23, 24]. However, when it comes to the estimation of disruption to transport and critical infrastructure capacities resulting from cyber attack, there appears to be a gap in the literature.

Disruption to air network infrastructure not only impacts the effective capacity of air networks but can propagate delay. Wu et al. [25] found that delays can be propagated due to reasons such as airport congestion, resource limitations, or even through connecting-flight delays. Wu et al. also noted that at least one airline in China had nearly 50% of its sequence flights suffering from such effects. Beatty et al. [26] came up with the concept of a delay multiplier to capture the amplification of an initial delayed flight through the day, estimating that the flight delay cost was more than 30 billion dollars every year. Other studies have found a range of flight delay propagation due to the increasing demand burden on the air transportation systems [27, 28, 29]. Modeling failure or disruption of critical infrastructure due to a natural or malicious act is generally well studied but is limited in both the specific exploration of transport infrastructure and in the use of graph- or agent-based techniques to assess the consequences of capacity loss across sectors. While there exist several studies looking at the disruption to specific critical infrastructure sectors utilizing system dynamics, agent-based, network, input-output, or high-level architecture models, only a single paper was identified that examined the disruption to the transportation sector [30]; and that study sought only to estimate the change in passenger demand in case of a physical attack on the US air infrastructure utilizing an input-output model. In this context our analysis contributes to the literature in two significant ways. First, our use of both a graph and an agent-based model are novel, providing both an approach to estimating capacity loss and propagation impacts in air transport. Second, the introduction of an index to measure effective capacity loss enables policy-makers to compare the impact of a range of cyber events more easily across any air network, and even serves as a means for measuring impacts across sectors.

3. MEASURING EFFECTIVE NETWORK CAPACITY LOSS IN AIR TRANSPORTATION

We measure effective capacity loss in an air network by imagining a set of airports as vertices connected by edges (flights) which ferry passengers between locations and where the loss of any airport or flight creates a loss in transport capacity (Figure 1).

Let us assume a weighted graph (G) below with the set of vertices (V={$v_1,v_2,\ldots,v_n$}), edge pairings *(E)*, and where W is a matrix of edge weights ($w_{i,j}$), where row i represents the individual vertices in G columns j equal the edge pairings between a specific vertex in row i and all other vertices in the graph, and whose value is equal to the number of passengers. For example, in Figure 1, $w_{1,2}$ and $w_{2,1}$ are equal to 5 while $w_{1,3}$ and $w_{3,1}$ are 0 as there is no connection between v_1 and v_3. Cyber events can have different impacts on services that support flight operations at an airport (T_{v_i}). These effects can range from slight delays to those that completely incapacitate operations (i.e. $0 \leq T_{v_i} \leq 1$). Examples might include a ransomware attack that disables tower communication systems ($T_{v_i} = 1$), a spear-phishing event that only compromises data, ($T_{v_i} = 0$), or a sustained DDoS event that degrades operations ($T_{v_i} = 0.3$).

To estimate an effective capacity loss of one or more airport vertices in the graph, we sum across all vertices in the graph, the operational effect of the attack per vertex (T_{v_i}), and multiply it by the sum of both the impact of the positional importance in the graph, measured by the eigenvector centrality of the vertex impacted by the cyber attack ($C_{v_i}^e$) over the sum of all the vertices' eigenvector centralities, and the volume of passengers traversing the affected vertex to other vertices divided by the sum of all passengers through the air network. Tuning parameters for both the positional importance (α) of the vertex as well as the volume of passengers the vertex supports (β) are included.

Therefore, for any graph G we can measure the Strategic Disruption Index (SDI) between 0 and 1:

FIGURE 1: A REPRESENTATIVE GRAPH (G) OF VERTICES AND EDGES
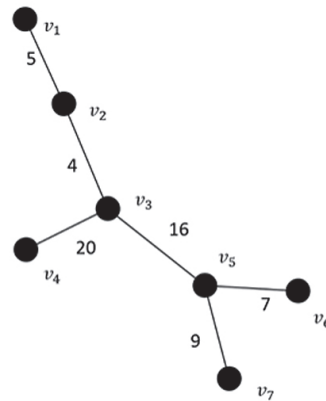
$$SDI = \sum_{v_i}^{n} \left(T_{v_i} \left[\alpha \left(\frac{C_{v_i}^e}{\sum_{v_i}^n C_{v_i}^e} \right) + \beta \left(\frac{\sum_{j_i}^n w_{i,j}}{\sum_i^n \sum_j^n w_{i,j}} \right) \right] \right)$$

Where: α + β = 1
$0 \leq T_{v_i} \leq 1$

For example, if we wanted to calculate the SDI for a ransomware attack on vertex v_3 tower communications ($T_{v_3} = 1$), we would take into account v_3's calculated eigenvector centrality (1) divided by the sum of all centralities in the graph (4.28) and the sum of edge weights connected to it (40) over the sum all edge weights (61). Assuming α and β are both equal to 0.5 (e.g., both factors are of equal importance), we can calculate an SDI of 0.45, indicating that the attack generated an effective capacity loss in this graph of 0.45 or 45% of its weighted capacity. Conversely, the same attack on v_1 would only generate an SDI of 0.08, or 8% of the effective weighted capacity of the entire air network. In this manner, we can use the SDI to differentiate the capacity loss across a range of airports, airlines, or supporting vendors for any number of cyber scenarios.

4. CONSTRUCTING AIR NETWORKS

We utilize year 2019 US Department of Transportation air carrier statistics to build graphs of airports and flight connections. Figure 2 highlights a national graph, with the top 10 airport codes identified by passenger volume and eigenvector centrality. We identified 374 airports with 5,481 connections that moved at least 1,000 passengers a month (Table I). The national air infrastructure can be described as a scale-free network with a few highly interconnected airports that stitch smaller regional locations together into a single air infrastructure. This follows a similar pattern found by Amaral et al. [19], with the degree distribution following a power law distribution.

FIGURE 2: GRAPH OF NATIONAL AIR TRANSPORT NETWORK
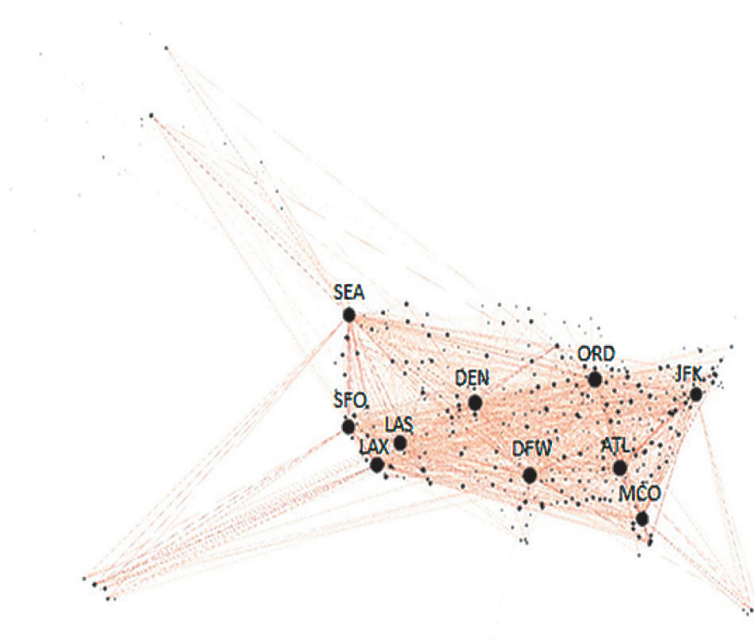
To assess the regional effects of disruption to air infrastructure, we divide the national air infrastructure into the 10 different Federal Emergency Management Agency (FEMA) regions (Figure 3). These regional graphs consist only of flights beginning and ending in the same region. Intra-regional networks vary not only in the number of vertices and passenger volume but also in structure. For example, FEMA Region 1 has relatively few airports (6) with an average degree centrality of 1.33, implying that most airports have few interconnections and that their structure is likely considerably different from other regions such as 4 and 9 that maintain both more vertices (61, 42) and greater average degree centrality (6.67, 6.79).

FIGURE 3: FEMA REGIONS

Figure 4 highlights two of the graph structures, demonstrating that there is a larger set of interconnections in Region 9 than in Region 1. We find that this diversity of network structure generates substantial differences in both the regional capacity loss and delay propagation effects from the disruption to either airport operations or airline infrastructure.

FIGURE 4: INTRA-REGIONAL AIR NETWORKS
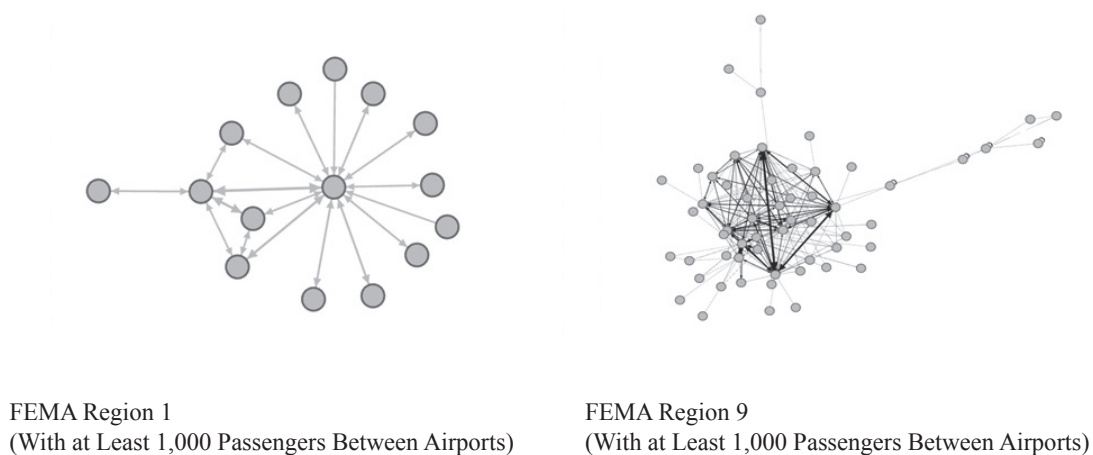
FEMA Region 1
(With at Least 1,000 Passengers Between Airports)

FEMA Region 9
(With at Least 1,000 Passengers Between Airports)

The differences in regional structures found in Table I, including in average degree, numbers of vertices, and edge connections, reflect the differences in geography and physical distance found in different parts of the country and which we find to drive differences in effective capacity loss and flight delay propagation.

TABLE I: AIR NETWORK DESCRIPTIVE STATISTICS

	Nodes	Edges	Diameter	Modularity	Avg. Degree
FEMA 1	6	8	3	0.32	1.33
FEMA 2	16	44	4	0.32	2.75
FEMA 3	17	54	3	0.21	3.18
FEMA 4	61	407	5	0.22	6.67
FEMA 5	45	181	3	0.18	4.02
FEMA 6	11	192	3	0.24	4.36
FEMA 7	5	8	2	0	1.60
FEMA 8	37	97	3	0.20	2.62
FEMA 9	42	285	3	0.18	6.79
FEMA 10	36	100	5	0.34	2.78
National	374	5481	7	0.25	14.66

5. ESTIMATING LOSS OF EFFECTIVE NETWORK CAPACITY IN US AIR INFRASTRUCTURE

Disruption to airport flight operations, including jetway functions, air traffic management systems, or even booking management systems, can have strategic impacts on the entire sector [32]. Using the SDI approach, our analysis shows that only a few well-connected airports (Table II) with large traffic volume generate the heaviest impacts; many small and regional airports have little overall impact on the national air network. These results broadly correlate with the scale-free structure of the national air network and confirm the experience of any air traveler who has been delayed when a weather event shuts down a major airport. For example, a cyber event that shuts down Hartsfield-Jackson Atlanta International Airport (ATL) generates an SDI of 0.04, or roughly 4% of the effective capacity of the national air transport system. Disruption at other major hubs, such as ORD, DEN, or DFW, induces similar

impacts to effective network capacity loss, whereas the loss of air operations at smaller airports such as in Boise, Idaho, represents less than 0.1% loss in effective national network capacity.

TABLE II: NATIONAL SDI BY AIRPORT

Airport	Strategic Disruption Index (SDI)
Hartsfield-Jackson Atlanta International (ATL)	0.04
O'Hare International (ORD)	0.03
Denver International (DEN)	0.03
Dallas/Ft. Worth International (DFW)	0.03
Los Angeles International (LAX)	0.03
Las Vegas McCarran International (LAS)	0.02
Charlotte Douglas International (CLT)	0.02
Seattle-Tacoma International (SEA)	0.02
Phoenix Sky Harbor International (PHX)	0.02
Orlando International (MCO)	0.02

Regional effects are also calculated utilizing each region's specific graph structure to estimate an SDI value. We find that while national SDI values remain largely consistent (at between 0.02 and 0.04) among the largest airports, disruption to the most important airports in each region can vary substantially (Table III). For example, in Region 9, which includes much of California, Arizona, and Nevada, the loss of LAX would constitute a 10% loss of the effective air capacity in the region. By contrast, the loss of St. Louis Lambert International Airport would be 42% of the regional network capacity. Similarly, an attack on Boston's Logan International would represent a loss of 30% of the regional network capacity.

TABLE III: REGIONAL AND NATIONAL SDI BY MOST AFFECTED AIRPORT IN FEMA REGION

FEMA Region	Top Disrupted Airport	Regional SDI	National SDI
1	Boston Logan International (BOS)	0.30	0.02
2	New York International (JFK)	0.20	0.01
3	Philadelphia International (PHL)	0.20	0.02
4	Hartsfield-Jackson International (ATL)	0.17	0.04
5	O'Hare International (ORD)	0.17	0.03
6	Dallas/Ft. Worth International (DFW)	0.16	0.03
7	St. Louis Lambert International (STL)	0.42	0.01
8	Denver International (DEN)	0.26	0.03
9	Los Angeles International (LAX)	0.10	0.03
10	Seattle-Tacoma International (SEA)	0.23	0.03

The range of values indicates that different network structures lead to substantial differences in impact. In some cases, where airports remain at the center of a regional hub-and-spoke network structure (e.g., Regions 1 and 3), the disruption of airport operations at a single location can generate impacts far exceeding that airport's influence in the national air infrastructure. This is primarily due to the small number of airports that are servicing regional flights and are highly reliant on a major airport (e.g., Boston's Logan International Airport). As seen in Figure 2, the network structure of Region 1 is tightly connected through a single vertex (BOS), yet Region 9 has more highly connected vertices limiting regional effective capacity loss from a single airport disruption.

How would attacks on airline infrastructure, including their vendor systems, compare with attacks on airports? Disruptions to some specific airline systems (e.g., disabling the ability to file a flight plan) can lead to the grounding of the entire air fleet across all airports they serve. In some cases, airlines provide much of the capacity at many airports, and thus attacks on them would disrupt large percentages of air capacity simultaneously across regions. Recent events at Delta in 2016, United Airlines in 2017, and Southwest in 2019 are representative. We find that attacks on an airline's air network generate significantly larger national effects (Table IV) than do attacks on a specific airport.

TABLE IV: NATIONAL SDI BY AIRLINE

Airline	Market Value	Vertices	Edges	Strategic Disruption Index (SDI)
Southwest	$126.45B	127	2505	0.10
Delta	$130.25B	215	3072	0.08
American Airlines	$131.59B	158	1953	0.08
United	$111.28B	181	2021	0.06
SkyWest	$21.39B	280	4084	0.02
Jet Blue	$41.44B	94	810	0.02
Alaska Airlines	$47.48B	114	771	0.02
Frontier	$22.57B	105	768	0.01
Hawaiian	$12.28B	30	104	0.01

The loss of capacity across potentially hundreds of air corridors simultaneously generates SDI values that are more than twice the impact of the largest and most central airports (such as ATL). This supports a general observation from scholars who point out that scale-free network structures are resilient after losing a single vertex but remain largely vulnerable to attacks on many highly connected nodes simultaneously [23]. For example, an attack against Southwest Airlines generates a 10% loss of effective network capacity across the United States, more than twice what was achieved in disrupting Atlanta's Hartsfield-Jackson International.

As airlines frequently manage operations using integrated services from third-party vendors, the loss of a single vendor's service can exacerbate the problem. Airlines that utilize the same vendor to provide critical services as part of their broad operations open the potential for a single third party to cause disruption to ground flights across *multiple airlines* simultaneously. For example, AeroData, a privately owned company providing flight inspection systems, suffered a system disruption in 2019 that forced major carriers United, Delta, Southwest, JetBlue, and Alaskan Airlines to cancel more than 7,000 flights throughout the day [31, 32]. The outage, while only lasting 40 minutes, would score a collective SDI value of 0.36, representing a capacity loss *nine times* greater than the loss of ATL operations. With national carriers responsible for the largest percentage of flights between major airports, the disruption of common vendor systems essential to flight operations presents the largest strategic impact to the effective network capacity of US air infrastructure. This type of attack highlights the challenge to scale-free network structures that, while resilient to the removal of

a single vertex, generate substantial impacts when several highly connected nodes are disrupted simultaneously. The use of a handful of key service vendors raises the possibility of single points of failure with the potential for widespread national air disruption extending far beyond the loss of flight operations at a single airport or airline.

6. MODELING PROPAGATION DELAYS

Disruption to air capacity also has the potential to propagate impacts through the entire air network. To measure this dynamic effect, we use an agent model leveraging flight-time information obtained from the Airline On-Time Performance Database, distributed by the Bureau of Transportation Statistics. Simulations are done in Python, with delays propagated to downstream flights using random sampling through Monte Carlo methods. While the SDI quantifies the weighted loss of an initial disruption on passenger air network capacity, our delay propagation captures the ensuing propagation of delays on the impending air flight network.

We define the propagation parameter, α, as the fraction of flights impacted by an initial delayed flight of the same airline, within a fixed-time interval (t_{dur}) from the scheduled time of arrival at the arrival airport. The agent-based algorithm is as follows:

1) x_0 flights are initially delayed from a cyber attack;
2) Each of the x_0 delayed flights impacts a fraction α of flights from the same airline within a certain time t_{dur} at the airport from which the flight is scheduled for departure;
3) Each of the subsequently delayed flights also causes delays at the corresponding airport at which they are scheduled to land, further propagating delays according to Rule 2, resulting in a cascade of delays through the day.

Previous work [29] found that α of between 0.02 (or 2% of downstream flights impacted by an initial delayed flight) and 0.25 (25% of downstream flights impacted by an initial delayed flight) reproduced clusters of propagating delayed flights in regular operating conditions. The higher α values reflected times in which there were more passengers leading to fewer buffers in the airline networks, such as during the holiday season. In our analysis we vary α as 0.02 or 0.25 and choose a fixed $t_{dur} = 1$ hour. For example, $\alpha = 0.25$ would mean that a delayed Delta flight that was supposed to land at 13:00 in Atlanta would impact 25% of subsequent Delta flights that are supposed to depart between 13:00 and 14:00. That delay would continue propagating across the Delta air network until the end of the day.

Figure 5 shows the cascading disruption in a scenario when an airport is shut down for an hour from 8:00–9:00 EST. Diagrams (a), (b), and (c) correspond to a disruption of the Atlanta Hartsfield-Jackson International Airport, at 9:00, 13:00, and 17:00 EST, respectively. Diagrams (d), (e), and (f) correspond to disruption of the Los Angeles International Airport during the same times as Diagrams (a), (b), and (c). The colors are a heatmap indicating the number of flights disrupted over the course of the day. Regions with greater disruption have more intense red colors, whereas regions of less disruption are colored blue. An attack on ATL airport propagates from East to West, whereas an attack that originates in LAX propagates from West to East.

FIGURE 5: AIRPORT CYBER ATTACK CASCADING DELAY SCENARIO ANALYSIS. DURATION OF THE ATTACK: 8:00–9:00 EST ON DECEMBER 1, 2019. TOP: ATL AIRPORT. BOTTOM: LAX AIRPORT.
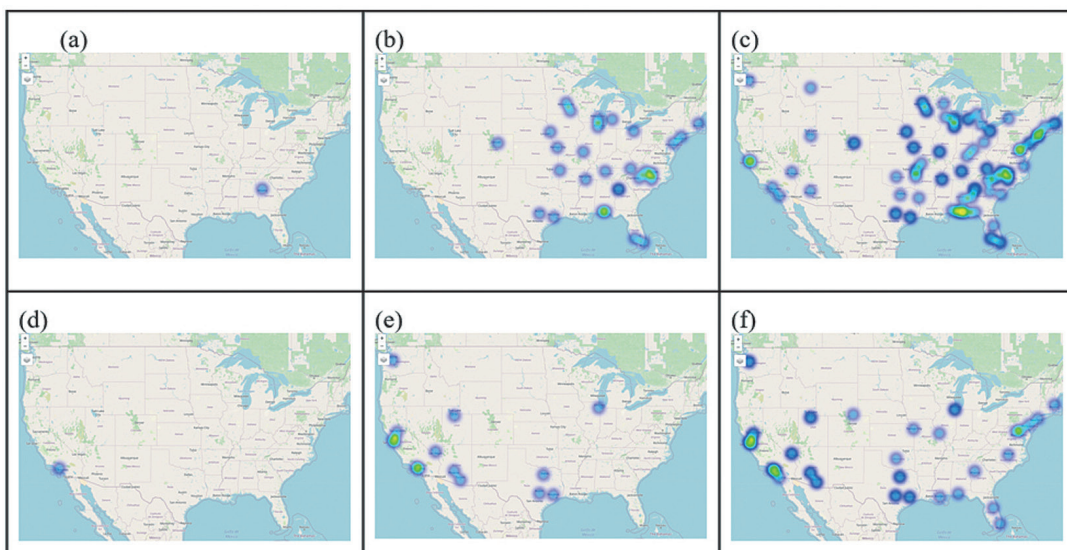
Figure 6 denotes the total fraction of delayed flights by the end of the day. Symbols denote attacks at 8:00–9:00 that impact the 10 most important airports in their respective FEMA region. Red lines are error bars from multiple simulations with the same α. The dashed black line is slope = 1, or national disruption = regional disruption. The low delay multiplier (α = 0.02) shows lower delay effects as compared to the larger delay multiplier (α = 0.25). Thus the propagation multiplier (α) acts as a "tuning" parameter to probe varying levels of cascade impacts.

FIGURE 6: EFFECTS OF DELAY MULTIPLIER ON FLIGHT DISRUPTION CASCADE.
(a) α = 0.02. (b) α = 0.25.
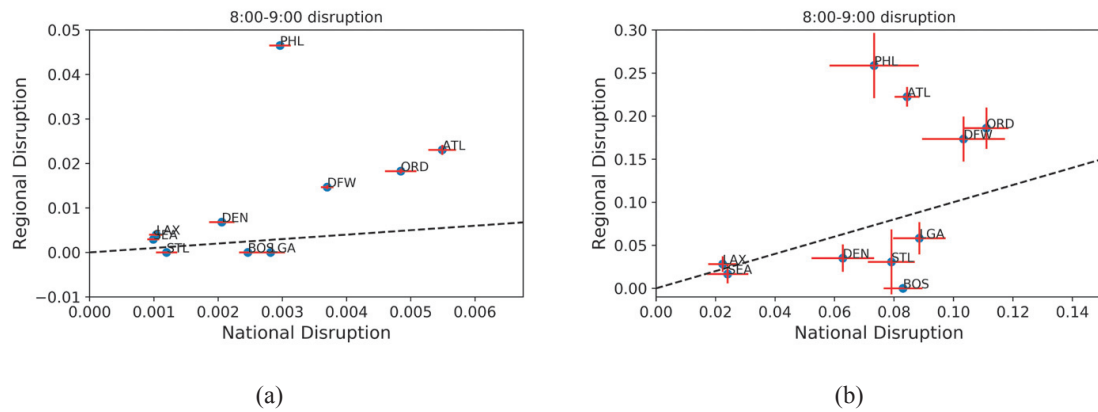
A smaller delay propagation parameter (α = 0.02) is associated with limited initial disruption between 8:00–9:00 EST, with hardly any cascade. Most airport disruptions are above the dashed line, showing that they cause larger regional disruptions than national ones. STL, BOS, and LGA are under the dashed line. STL and BOS cause little to no regional disruption due to the lack of intra-regional flights *by major airlines* in FEMA Regions 1 (no regional flights on December 1) or 7 (13 regional flights on December 1), whose flights are well distributed through the day. By contrast, FEMA Region 3's reliance on larger carriers is significantly impacted by the disruption of PHL, creating the largest regional delay disruption. This is likely due to the large presence of American Airlines in FEMA Region 3. It should be noted that the Airline On-Time Performance Database does not contain information on small regional carriers, which might lead to the underestimation of delay propagations in regions that are have a larger dependence on small regional carriers, such as FEMA Region 1. National delay impacts largely align with the large hubs, with ATL, ORD, and DFW accounting for the largest national disruptions.

Larger delay parameters (α = 0.25) are associated with larger downstream propagation. PHL, ATL, and ORD generated the largest regional disruptions; ORD, DFW, and LGA generated the largest national disruptions; and West Coast airports DEN, LAX, and SEA saw smaller disruptions, likely due to the early hour of the event (e.g., 6:00 MST/5:00 PST). Changes in the timing of an attack appear to also induce differences in national and regional delay propagation. In Figure 7b, we see an overall shift of disruption to lower national impacts when the cyber attack is later in the day, as the originating delay has less time to propagate to other connections during the day. However, cyber attacks on West Coast airports, including DEN, LAX, and SEA, respectively, lead to larger disruptions when the attack originates later in the day. Cascading delay disruptions are found to be sensitive to both geographical and temporal variations.

FIGURE 7: TIME OF DAY IMPACT ON CASCADE. (A) DELAY CASCADE IN A SCENARIO WHERE THE RESPECTIVE AIRPORTS ARE DISRUPTED FROM 12:00–13:00 EST. (B) RELATIVE DISRUPTION IN A 12:00–13:00 SHUTDOWN VS. 8:00–9:00 SHUTDOWN.

(a)   (b)
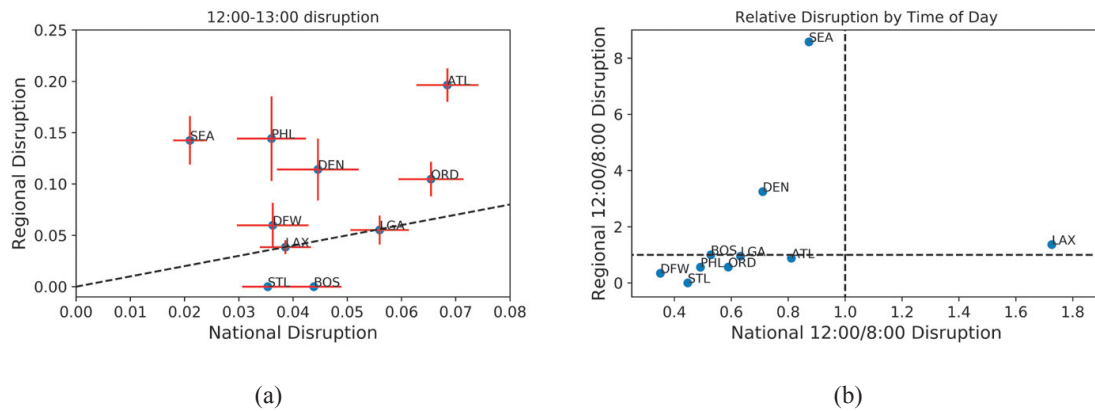
Attacks on airline infrastructure also cause delays, as aircraft are unable to make connections when proprietary systems do not allow for normal flight operations. We study a scenario in which flights are disrupted on December 1, 2019, corresponding to the five largest carriers by flight volume: American (AA), Delta (DL), United (UA), Southwest (WN), and Alaskan (AS) Airlines, respectively. We find that the five largest airline network structures vary in the level of both national disruption and regional impacts corresponding to the locations of its hub operations. Figure 8 highlights both the national and regional propagation delays across the five largest carriers. In the most extreme case, an attack that disables operations at Alaskan Airlines impacts almost exclusively a single region (Region 10). Figure 9 highlights the range of national and regional effects to compare the impacts of airline shutdowns, plotting the fraction of flights. Here the numbers for the air carriers correspond to the FEMA region that is most impacted (in terms of the fraction of delayed flights). For example, during a DL shutdown from 12:00–13:00, 15% of flights are delayed across the country through the day, and in FEMA Region 4, 30% of flights are delayed. In general, we find that, similarly to the analysis of effective capacity loss, delay disruptions from attacks on airline carriers are larger than those from attacks on airports.

FIGURE 8: IMPACTS OF AIRLINE SHUTDOWNS FROM 12:00–13:00 ON DECEMBER 1, 2019. LEFT: PERCENTAGE OF TOTAL FLIGHTS DELAYED CASCADING THROUGH THE DAY NATIONALLY AND BY FEMA REGION (AS INDICATED IN THE LEGEND). RIGHT: SNAPSHOTS OF DELAY AT THE END OF THE DAY.
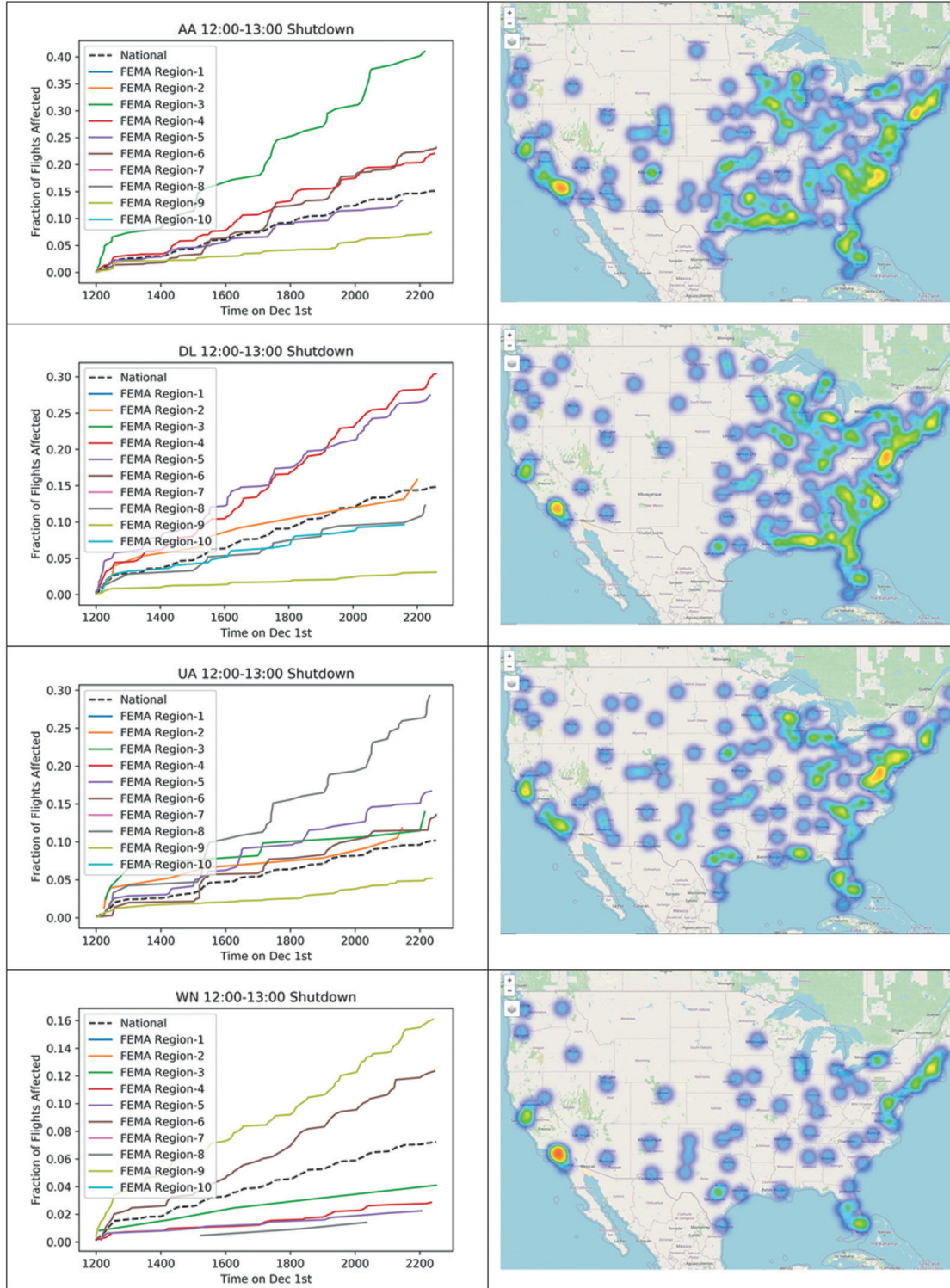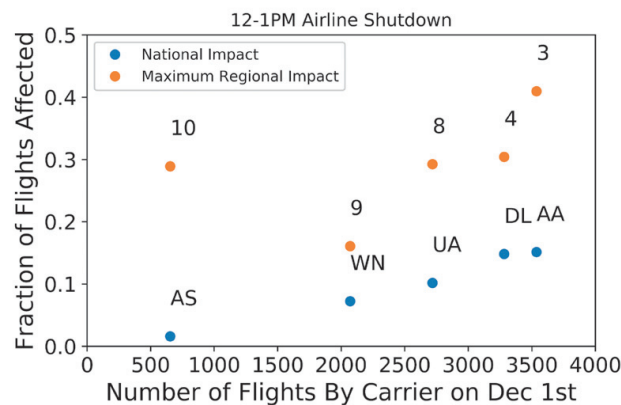
Additionally, the four regions that endure the largest disruptions from airline attacks are the same four regions that have the biggest disruptions as a result of airport attack (Regions 3, 4, 8, and 10). FEMA regions are particularly susceptible to strategic effects due to the combination of regional reliance on the dominant airport and the same airport being the hub of a large national carrier. Regions where carrier hub airports are located are more vulnerable to cascading delays that originate either from carrier or airport shutdown, due to the interdependencies between airport and carrier network structures. By contrast, national impacts are not dependent on the structure of airline networks, due to the national redundancy of airline hubs in multiple airports across regions. While we assumed a uniform delay propagation factor ($\alpha = 0.25$) across airline networks, studies have indicated that certain point-to-point carriers have lower risks for cascading delays [33, 34].

7. CONCLUSION

The growing threat of attacks against critical infrastructure is an enduring concern among national policy-makers. Key to managing the risk of these threats is the ability to effectively measure the strategic effect a cyber attack has on the series of interconnected organizations. Organizational dependencies, including third-party vendors, can create disparate and complex impacts to capacity and delay propagation. These complex sets of interdependencies challenge the ability of policy-makers to effectively prioritize defensive resources.

National disruptive effects propagated primarily through airlines highlight the potential for substantial impacts through the degradation of third-party vendors who provide service across multiple airlines. The combination of capacity loss and propagation

delay across major hubs simultaneously serves as an effective attack on the scale-free structure of the national air infrastructure. While operators utilize these vendors to take advantage of efficiencies, they add critical vulnerabilities to the whole of the transport sector. Attacks on a single operational system (e.g., passenger booking) have the potential to disrupt air infrastructure more than attacks on entire airports. Threats to vendors who provide for terminal management, passenger facilitation, airside operations, and information management all represent avenues of strategic disruption not presently accounted for in the effects literature. Decisions regarding design, development, and deployment of key flight operational systems, all made by *private actors*, can generate substantial *public risk*.

Policy-makers should examine the role that service providers play to better assess the risk these services present to US air infrastructure. For example, the temporary disabling of the AeroData service highlights how the tight coupling of vendors with essential flight operations by airlines creates the potential for pervasive and highly disruptive impacts. The disruption to flight operations generated an effective capacity loss, albeit for 40 minutes, of over 36% (SDI = 0.36) of the nation's capacity with a 200–300% increase in flight delays (Figure 10). Russian dispersal of the NotPeyta ransomware through M.E.DoC software in Ukraine and, more recently, the attack on Solarwinds further highlights the strategic risk to organizations using third-party vendors.

FIGURE 10: CASCADING DELAYS ON APRIL 1, 2019, DUE TO THE AERODATA SERVICE OUTAGE WHICH IMPACTED MULTIPLE AIRLINES, COMPARED WITH DELAYS DURING THE FOLLOWING DAYS
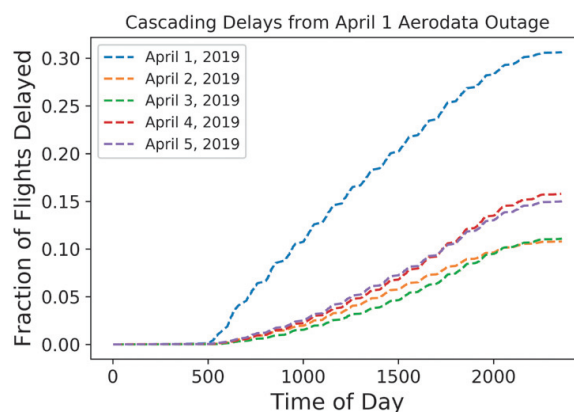
Significant *regional* effects on effective network capacity can also occur if specific airport or airline infrastructure is targeted. In some cases, regional networks have concentrated connections in a few airports, leading to substantial disruption to regional flights when a single airport is made inoperable. Regions with greater numbers of connected airports maintain greater resilience in their air sector, whereas

regions with a dominant metropolitan area (e.g., Boston) will be more susceptible to regional disruption from the loss of a single airport or large airline servicing intraregional flights. Furthermore, disruption to low-resilience airport infrastructure or to airline capacity heavily concentrated in those locations can lead to substantial regional delay propagation in addition to the significant loss of effective capacity. Estimations of risk at the state and local level might vary, given this difference from the national results; these variations highlight the general degree of nuance one needs to employ when estimating the collective impact to effective network capacity. While several qualitative [35, 36] and quantitative studies [37] explore interconnected vulnerabilities in critical infrastructure, the complexity of interdependence creates substantial challenges to measuring strategic impacts [38]. The approach explored in this paper expands on prior efforts to create a more extensible method for comparing strategic effects between sectors.

Our framework complements existing approaches through the combination of network analysis and computing network flows. The advantages of this approach are the relative ease and adaptability to various other infrastructures, such as other transportation networks, power supply networks, and water networks. For example, the 2003 Northeastern blackout features a short-time cascade, which can be modeled using a similar approach.

A risk-based approach to cyber security defense is at the heart of the US public efforts to promote resilience in critical infrastructure. While both the US 2018 National Cybersecurity and relevant executive action define and promote defense of the nation's 16 critical infrastructures as an essential element of the defensive strategy, the ability to quantify the range of strategic effect remains an ongoing challenge in the field. Policy-makers who are charged with developing means of measuring national risk can apply the approach in this paper to assess the interdependence of organizations and prioritize resources to best limit the range of strategic risk to any number of critical infrastructures.

REFERENCES

[1]   E. Kenneally, L. Randazzese, and D. Balenson, "Cyber risk economics capability gaps research strategy," in *2018 Int. Conf. Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*, 2018, pp. 1–6.

[2]   C. Harry, "The challenge of assessing strategic cyber security risk in organizations and critical infrastructure," *Cyber Security: A Peer-Reviewed J.* vol. 4, pp. 58–69, 2020.

[3]   Government Accountability Office, "Commercial Aviation: Information on Airline IT Outages," GAO-19-514, 2019.

[4]   S. Ikeda. "Impact of cyber attacks on RavnAir more damaging than first thought; flights may be grounded for a month." https://www.cpomagazine.com/cyber-security/impact-of-cyber-attacks-on-ravnair-more-damaging-than-first-thought-flights-may-be-grounded-for-a-month/ (accessed Dec. 21, 2020).

[5]   Reuters. "Hackers ground 1,400 passengers at Warsaw in attack on airline's computers." http://www.theguardian.com/business/2015/jun/21/hackers-1400-passengers-warsaw-lot (accessed Dec. 21, 2020).

[6]   C. Cimpanu. "Swedish air space shut down by cyber-attacks, officials blame Russia." news.softpedia.com/news/swedish-air-space-shut-down-by-cyber-attacks-officials-blame-russia-502919.shtml (accessed Dec. 21, 2020).

[7]   P. Polityuk and J. Stubbs. "New wave of cyber attacks hits Russia, other nations." Reuters. https://www.reuters.com/article/us-ukraine-cyber/new-wave-of-cyber-attacks-hits-russia-other-nations-idUSKBN1CT21F (accessed Dec. 21, 2020).

[8]   C. Simmons, C. Ellis, S. Shiva, D. Dasgupta, and Q. Wu, "AVOIDIT: A cyber attack taxonomy," *9th Annu. Symp. Inf. Assurance*, pp. 2–12, 2014.

[9]   I. Agrafiotis, J. R. C. Nurse, M. Goldsmith, S. Creese, and D. Upton, "A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate," *J. Cybersecurity*, vol. 4, pp. 1–15, 2018.

10]   A. Cresswell and S. Hassan, "Organizational impacts of cyber security provisions: A sociotechnical framework," in *2007 40th Annu. Hawaii Int. Conf. Syst. Sci. (HICSS'07)*, pp. 98-125 2007.

[11]   C. Harry and N. Gallagher, "Classifying cyber events," *J. Inf. Warfare*, vol. 17, pp. 17–31, 2018.

[12]   S. Vivek, D. Yanni, P. J. Yunker, and J. L. Silverberg, "Cyberphysical risks of hacked internet-connected vehicles," *Physical Rev. E*, vol. 100, p. 012316, 2019.

[13]   D. Kundur, X. Feng, S. Liu, T. Zourntos, and K. L. Butler-Purry, "Towards a framework for cyber attack impact analysis of the electric smart grid," in *2010 First IEEE Int. Conf. Smart Grid Commun.*, 2010, pp. 244–249.

[14]   R. Taormina, S. Galelli, H. C. Douglas, N. O. Tippenhauer, E. Salomons, and A. Ostfeld, "A toolbox for assessing the impacts of cyber-physical attacks on water distribution systems," *Environmental Modelling & Software*, vol. 112, pp. 46–51, 2019.

[15]   J. Petit and S. E. Shladover, "Potential cyberattacks on automated vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, pp. 546–556, 2014.

[16]   R. Dieye, A. Bounfour, A. Ozaygen, and N. Kammoun, "Estimates of the macroeconomic costs of cyber-attacks," *Risk Manage. and Insurance Rev.*, vol. 23, pp. 183–208, 2020.

[17]   J. R. Santos, Y. Y. Haimes, and C. Lian, "A framework for linking cybersecurity metrics to the modeling of macroeconomic interdependencies," *Risk Anal.*, vol. 27, pp. 1283–1297, 2007.

[18]   D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, pp. 440–442, 1998.

[19]   L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, "Classes of small-world networks," *Proc. Nat. Acad. Sci.*, vol. 97, pp. 11149–11152, 2000.

[20]   V. Kalapala, V. Sanwalani, A. Clauset, and C. Moore, "Scale invariance in road networks," *Physical Rev. E*, vol. 73, no. 2, p. 026130, 2006.

[21]   P. Sen, S. Dasgupta, A. Chatterjee, P. A. Sreeram, G. Mukherjee, and S. S. Manna, "Small-world properties of the Indian railway network," *Physical Rev. E*, vol. 67, pp. 036106, 2003.

[22]   P. S. Dodds and D. H. Rothman, "Geometry of river networks. III. Characterization of component connectivity," *Physical Rev. E*, vol. 63, p. 016117, 2000.

[23]   R. Albert, H. Jeong, and A. L. Barabasi, "Error and attack tolerance of complex networks," *Nature*, vol. 406, pp. 378–382, 2000.

[24]   A. Bonato, "A course on the web graph," *American Mathematical Soc.*, vol. 89, 2008.

[25]   W. Wu, H. Zhang, T. Feng, and F. Witlox, "A network modelling approach to flight delay propagation: Some empirical evidence from China," *Sustainability*, vol. 11, p. 4408, 2019.

[26]   R. Beatty, R. Hsu, L. Berry, and J. Rome, "Preliminary evaluation of flight delay propagation through an airline schedule," *Air Traffic Control Quart.*, vol. 7, pp. 259–270, 1999.

[27]   N. Kafle and B. Zou, "Modeling flight delay propagation: A new analytical-econometric approach," *Transp. Res. Part B: Methodological*, vol. 93, pp. 520–542, 2016.

[28]   N. Pyrgiotis, K. M. Malone, and A. Odoni, "Modelling delay propagation within an airport network," *Transp. Res. Part C: Emerging Technologies*, vol. 27, pp. 60–75, 2013.

[29]   P. Fleurquin, J. J. Ramasco, and V. M. Eguiluz, "Systemic delay propagation in the US airport network," *Sci. Reps.*, vol. 3, p. 1159, 2013.

[30]   M. Iturriza, L. Labaka, J. M. Sarriegi, and J. Hernantes, "Modelling methodologies for analysing critical infrastructures," *J. Simulation* vol. 12, no. 2, pp. 128–143, 2018.

[31]   Deutsche Welle. "AeroData software outage delays hundreds of US regional flights." https://www.dw.com/en/aerodata-software-outage-delays-hundreds-of-us-regional-flights/a-48152813 (accessed Dec. 21, 2020).

[32]   D. Spaniel and P. Eftekhari, "Hacking our nation's airports: Cyber-kinetic threats to the technologies running airport operations," *Inst. Crit. Infrastructure Technol.*, 2019.

[33] Z. Zgodavova, R. Rozenberg, and S. Szabo, "Analysis of point-to-point versus hub-and-spoke airline networks," in *2018 XIII Int. Sci. Conf. – New Trends in Aviation Development (NTAD)*, 2018, pp. 158–163.

[34] G. Cook and J. Goodwin, "Airline networks: A comparison of hub-and-spoke and point-to-point systems," *J. Aviation/Aerospace Educ. & Res*., vol. 17, pp. 2, 2008.

[35] N.J. Carhart and G. Rosenberg, "A framework for characterising infrastructure interdependencies," *Int. J. Complexity in Appl. Sci. and Technol*., vol. 1, pp. 35–60, 2016.

[36] A. Laugé, J. Hernantes, and J. M. Sarriegi, "Critical infrastructure dependencies: A holistic, dynamic and quantitative approach," *Int. J. Critical Infrastructure Protection*, vol. 8, pp. 16–23, 2015.

[37] M. Iturriza, L. Labaka, J. M. Sarriegi, and J. Hernantes, "Modelling methodologies for analysing critical infrastructures," *J. Simulation*, vol. 12, 128–143, 2018.

[38] C. Nan, I. Eusgeld, and W. Kroger, "Analyzing vulnerabilities between SCADA system and SUC due to interdependencies," *Rel. Eng. and Syst. Safety*, vol. 113, 76–93, 2013.

Adversary Targeting of Civilian Telecommunications Infrastructure

Keir Giles
Conflict Studies Research Centre
Northamptonshire, United Kingdom
keir.giles@conflictstudies.org.uk

Kim Hartmann
Conflict Studies Research Centre
Northamptonshire, United Kingdom
kim.hartmann@conflictstudies.org.uk

Abstract: The response to the pandemic by states, organisations, and individuals in 2020 highlighted critical dependency on communications systems underpinned by cyber infrastructure. Without the benefits of connectivity, governments would have faced greater challenges governing, societies would have found it even harder to maintain cohesion, more companies would have ceased to operate altogether, and personal isolation would have been a vastly more difficult experience.

And yet, it is precisely this connectivity within and between NATO states that some adversaries are preparing to attack in time of conflict, including through physical or kinetic means. Russia in particular has long invested in probing vulnerabilities of civilian internet and telecommunications infrastructure, and this programme was urgently ramped up to unprecedented levels of intensity after the seizure of Crimea in 2014 demonstrated the power of total information dominance achieved through targeting critical information assets.

Besides Russia, China and a number of other states are also rapidly developing counter-space capabilities that would pose a direct threat to critical civilian communications services. This has obvious implications for crisis management even before overt state-on-state conflict. Vulnerabilities have been sought in all domains: maritime (subsea cables), space (communications satellites), land (fibre optic nodes), and online (targeting specific media sources for neutralisation). The VPNFilter malware exposed in mid-2018, in addition to its cybercrime or cyber-espionage capabilities, demonstrated the ambition to render large numbers of ordinary users in NATO countries simply unable to communicate.

Recognising and responding to this emerging disruptive threat and its potential human, societal, and state impact is critical to the defence of NATO states – still more so in the case of disruption to normal life by events such as the pandemic. The threat to cyber-

physical systems not ordinarily considered a military target must be recognised, and their defence and security prioritised. This paper outlines the threat and recommends a range of mitigation strategies and measures.

1. INTRODUCTION

"In the modern era you can achieve the same effect as used to be achieved in, say, World War Two by bombing the London docks or taking out a power station, by going after the physical infrastructure of cyberspace."

Mark Sedwill, former National Security Adviser, UK Cabinet Office [1]

On Christmas Day 2020, a suicide vehicle-borne improvised explosive device (SVBIED) detonated in central Nashville, Tennessee, next to a facility operated by telecoms provider AT&T.[2] The incident "brought communications in the region, from Georgia to Kentucky, to a halt, affecting 911 call centers, hospitals, the Nashville airport, government offices and individual mobile users… businesses big and small". The extent of the communications failures and subsequent disruption demonstrated not only that the AT&T facility represented a single point of failure for telecommunications networks across an extensive area of the United States, but also that local and regional government offices and essential services had no fallback options for maintaining communications.[3]

The Nashville attack is considered an isolated incident, carried out by a single troubled individual. But the vulnerability and lack of resilience demonstrated by this one event will have been of intense interest to nation states that wish harm to the United States and its allies, and in particular, to those that in time of conflict aim to target critical information infrastructure and the connectivity it provides. The disruption caused by one attack would be substantially increased by a simultaneous, coordinated campaign

---

1   "Joint Committee on the National Security Strategy", 18 December 2017, http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/national-security-strategy-committee/work-of-the-national-security-adviser/oral/75927.pdf

2   Kimberlee Kruesi, Michael Balsamo, and Eric Tucker, "Downtown Nashville Explosion Knocks Communications Offline", AP, 25 December 2020, https://apnews.com/article/Nashville-explosion-Christmas-52708bfd05e4f6ff433cc404443c65d4

3   Yihyun Jeong and Natalie Allison, "Nashville Bombing Exposed 'Achilles Heel' in Area Communications Network", *Nashville Tennessean*, 29 December 2020, https://www.tennessean.com/story/news/local/2020/12/29/nashville-bombing-area-communications-network-exposed-achilles-heel/4062089001/

against key information nodes, presenting a serious challenge to governance and the normal functioning of society in the victim state.

This paper considers the challenge of direct intervention against physical infrastructure in the context of a cyber, information, or conventional conflict. It reviews the stated or implicit ambition of adversaries to achieve information dominance, disruption, or destruction through action against civilian telecommunications infrastructure during or before overt conflict, and the implications for NATO nations. Sections review adversary activities toward this end in different domains: space, subsea, on land, and online. The paper then concludes with a set of proposed means of mitigating a range of vulnerabilities.

A. Dependence on Connectivity

The response to the pandemic by states, organisations, and individuals in early 2020 highlighted the critical dependency of societies on communications systems underpinned by cyber infrastructure. Without the benefits of hyperconnectivity, governments would have faced greater challenges governing, communities would have found it even harder to maintain cohesion, more private sector companies would have ceased to operate altogether, and personal isolation would have been a vastly more difficult and unpleasant experience. But even in normal times, dependence on always-on internet and telecommunications in many states has grown to the point that their denial or interdiction would cause severe challenges.

A number of distinct phenomena exacerbate this problem. First, the assumption that internet access is a normal default state leads to a neglect of redundancy and resilience, such that when, for example, Google services are briefly unavailable, large numbers of organisations find their normal business entirely paralysed.[4] Second, the ongoing rollout of the internet of things (IoT) has at times been accompanied by insufficient consideration of fallback modes for communications outages. This means that when backbone providers such as Amazon Web Services are disrupted, the impact is not only on commerce, logistics, media outlets and governance, but also on families finding their smart homes and smart devices have stopped working.[5] Furthermore, malign actors taking remote control of connected devices with no failsafes will increasingly present severe challenges to everyday activities.[6] Third, on an individual level,

---

4   "Google Cloud Infrastructure Components Incident #20013", Google, 14 December 2020, https://status. cloud.google.com/incident/zall/20013; Erika Varagouli, "'Google Down': How Users Experienced Google's Major Outage", Semrush, 15 December 2020, https://www.semrush.com/blog/google-down-how-users-experienced-google-major-outage/

5   Jay Greene, "Amazon Web Services outage hobbles businesses", *Washington Post*, 25 November 2020, https://www.washingtonpost.com/business/economy/amazon-web-services-outage-stymies-businesses/2020/11/25/b54a6106-2f4f-11eb-860d-f7999599cbc2_story.html; "AWS: Amazon web outage breaks vacuums and doorbells", BBC, 26 November 2020, https://www.bbc.co.uk/news/technology-55087054

6   Lorenzo Franceschi-Bicchierai, "We Spoke to a Guy Who Got His Dick Locked in a Cage by a Hacker", *Vice*, 28 January 2021, https://www.vice.com/en/article/4ad5xp/we-spoke-to-a-guy-who-got-his-dick-locked-in-a-cage-by-a-hacker

reliance on connected devices has led to an atrophy of skills required for when they are not available. This has been primarily highlighted to date in one of the clearest and simplest examples: the growing inability not only among the general public but even among military recruits to read maps and thus be able to navigate when disconnected.[7]

Until now, the incidents that demonstrate these vulnerabilities have been isolated, brief, and the result of technical errors or natural incidents rather than deliberate attack – but as with the Nashville blast, they give an indication of the potential damage if the reverse were true. The impact of this kind of attack would vary between NATO nations and even within them: the vulnerability will be even greater in countries with high degrees of connectivity and extensive adoption of near-universal online government and financial services, such as Estonia, than in countries that are relatively backward in this regard, such as the United States.[8] But in all cases, both the opportunities for carrying out this kind of attack and its probable impact are greatly increased if the attractiveness of civilian telecommunications infrastructure as a target for adversaries is underestimated.

B. Russia

The state that has most clearly acted on this attractiveness is Russia. The underlying principles of attacks on communications nodes are neither unique nor new, but it is primarily Russia that has both demonstrated and learned from the value in modern conflict of kinetic attacks that facilitate information outcomes, as opposed to the reverse.

Recent shifts in Russian thinking about the potential power of information warfare go to the heart of how wars are won: whether by destroying the enemy, or by rendering the enemy unable to fight.[9] For the latter purpose, the use of information operations against adversary populations and societies is part of an unbroken tradition in the institutional culture of Russia's military, intelligence, and political leadership that reaches back not only into Communist times but even before.[10] This includes information interdiction. In the current century this has been exercised via the internet: the socio-cyber attacks on Estonia in 2007 included crude attempts at cutting communications between government and citizens and with the outside world, modified and implemented with greater success against Georgian government communications the following

---

7   Danielle Sheridan, "Soldiers Must Know How to Read Maps Because Satellites Could Be Lost, Commander Field Army Says", *Daily Telegraph*, 4 December 2020, https://www.telegraph.co.uk/news/2020/12/04/soldiers-must-know-read-maps-satellites-could-lost-commander/

8   Olga Khazan, "America's Terrible Internet Is Making Quarantine Worse", *Atlantic*, 17 August 2020, https://www.theatlantic.com/technology/archive/2020/08/virtual-learning-when-you-dont-have-internet/615322/

9   Keir Giles, "Russia's 'New' Tools for Confronting the West: Continuity and Innovation in Moscow's Exercise of Power", Chatham House, March 2016, https://www.chathamhouse.org/sites/default/files/publications/2016-03-russia-new-tools-giles.pdf

10   Bilyana Lilly and Joe Cheravitch, "The Past, Present, and Future of Russia's Cyber Strategy and Forces", 2020 12th International Conference on Cyber Conflict, May 2020, https://www.ccdcoe.org/uploads/2020/05/CyCon_2020_8_Lilly_Cheravitch.pdf

year.[11] But the prehistory of this kind of operation includes the traditional seizure or destruction of civilian broadcast facilities and telephone and telegraph exchanges at the first stage of any attempt at regime change, whether imposed from abroad or by notionally domestic actors – as exemplified by a previous Moscow-backed attack on Estonia, the attempted coup in 1924.[12] Similarly, during the Cold War, part of the mission of KGB and GRU (Main Intelligence Directorate) sabotage teams inserted into Western countries was to seize or destroy communications and radio and TV broadcasting facilities.[13]

The extension of this principle into targeting internet infrastructure had been flagged in Russian conceptual writing on information warfare. An authoritative analysis of the new capabilities required by Russia following the armed conflict in Georgia in 2008 noted that "it is necessary to develop a centre for the determination of critically important information entities of the enemy, including how to eliminate them physically".[14] As in other cases, realisation of the offensive potential of operations of this kind was accompanied, or perhaps driven, by recognition of Russia's own previous vulnerability in this regard. The security and intelligence agencies' calls for greater attention to information security were in part founded on the concerns that "destruction and disorganisation of information infrastructure… on the scale of weapons of mass destruction is possible".[15]

But it was the seizure of Crimea in 2014 that provided a case study of information dominance facilitating an almost bloodless geopolitical gain, and consequently gave substantial impetus to Russia's interest in the potential vulnerabilities of NATO allies' civilian communications infrastructure. After gradually establishing control over traditional media in the days leading up to the operation to take the peninsula, Russian troops took over the Simferopol Internet Exchange Point and telecommunications cable connections to the mainland.[16] Together these operations gave Russia complete control of the Crimean information space, isolating it from the outside world.[17] The result was public perception of events in Crimea being determined exclusively by

11 Sean Ainsworth, "The Evolution of the Russian Way of Informatsionnaya Voyna", in Reuben Steff et al. (editors), *Emerging Technologies and International Security: Machines, the State, and War* (Routledge, 2020), pp. 137–152.
12 Merle Maigre, "Nothing New in Hybrid Warfare: The Estonian Experience and Recommendations for NATO", German Marshall Fund of the United States (GMF) Policy Brief, February 2015, p. 2.
13 "The Soviet Army: Specialized Warfare and Rear Area Support", FM 100-2-2, US Army, 16 July 1984, p. 5-4.
14 "Russia is Underestimating Information Resources and Losing out to the West", *Novyy Region*, 29 October 2008.
15 Vladimir Markomenko, "Невидимая затяжная война" (Invisible protracted war), *Nezavisimoye voyennoye obozreniye*, No. 30, 16 August 1997.
16 "Кримські регіональні підрозділи ПАТ 'Укртелеком' офіційно повідомляють про блокування невідомими декількох вузлів зв'язку на півострові" (Crimean regional divisions of PJSC "Ukrtelecom" officially report the blocking of several communication nodes on the peninsula by unknown persons), Ukrtelekom, 28 February 2014, https://www.scm.com.ua/news/ukrtelecom-s-statement
17 Shane Harris, "Hack Attack. Russia's First Targets in Ukraine: Its Cell Phones and Internet Lines", *Foreign Policy*, 3 March 2014, http://foreignpolicy.com/2014/03/03/hack-attack/

Russia, which contributed greatly to preventing resistance to the takeover by the civilian population.

The operation showed that advanced cyber capabilities are not necessary to achieve total control of an internet and telecommunications network if it is possible to mount a physical intervention against network infrastructure, the reverse of the more commonly considered scenario where cyber vulnerabilities are exploited for damaging physical effect.[18] This recognition appears to lie behind an intense and urgent subsequent pattern of activity by Russian military and intelligence organisations directed at civilian internet and telecommunications facilities across multiple continents. The end goal may be to interdict information through use of cyber, electronic warfare (EW), or kinetic activity, denying NATO governments the ability to communicate with their citizens in time of conflict and denying populations access to outside information, in an attempt to replicate the success delivered by total information dominance in Crimea. But even if Russia's objectives are limited to the military aims of denying, disrupting, or degrading NATO's ability to communicate, navigate, and target opposing forces, attempts to do so through destructive intervention against internet infrastructure would have profound second- and third-order effects on civil society during even a brief confrontation.

The remainder of this paper therefore considers the various domains in which threats to the infrastructure underpinning civilian internet and telecommunications services arise: subsea, in space, on land (including by electronic warfare), and in cyber and information space. Throughout, it should be remembered that the nature of the threat will vary between adversaries, because not all adversaries are identical and they will play to their strengths; for instance, the potential abuse of hardware and firmware dominance by China in Western telecommunications networks is an enduring source of concern, but Russia does not have the kind of ICT (information and communications technology) sector that would allow it to use a comparable vector of attack. Specifically considering preparation for physical interventions against civilian infrastructure, although Russia is not the only state with apparent ambitions of this kind, it is Russian actions that are by far the most widely reported. It is probably not possible to determine from open sources why this is so – whether other countries attach lesser importance to mapping the infrastructure of their potential adversary in this way, or whether, conversely, they ascribe greater importance to doing so in a manner that remains undetected.

---

18   Owen Matthews, "Russia's Greatest Weapon May Be Its Hackers", *Newsweek*, 5 July 2015, http://www.
      newsweek.com/2015/05/15/russias-greatest-weapon-may-be-its-hackers-328864.html

2. INFORMATION INTERDICTION

A. Subsea

In the period after the seizure of Crimea, Russia appeared to prioritise other concerns, such as speed, over remaining unobserved. This was especially apparent in the case of the first Russian activities that came to widespread public notice, namely investigation of subsea communications cables for either intelligence exploitation or disruption. The Russian agency primarily responsible for this, the *Glavnoye upravleniye glubokovodnykh issledovaniy* (Main Directorate for Deep-Water Research, GUGI), is a highly secretive organisation that until 2014 operated with such stealth that its purpose, and even its existence, very rarely appeared in open sources.[19] After Crimea, however, the apparent urgency of the task meant GUGI and its vessels attracted sufficient attention that they routinely featured in public reporting in the West.[20]

Concern rose that Russia was seeking the ability to choke off vital international communication channels at will, a task made easier by the fact that the majority of subsea cables are privately owned and their locations publicly known. Submarine cables carrying data, and in some cases those carrying power, present critical vulnerabilities to destructive intervention, with the potential for enormously damaging economic as well as societal disruption.[21] Targeting them would meet a wide range of Russian objectives; according to former SACEUR (Supreme Allied Commander Europe) Jim Stavridis, these would include "a rich trove of intelligence, a potential major disruption to an enemy's economy and a symbolic chest thump for the Russian Navy".[22] While the problem is potentially global in scope, Russian activities around the continental United States, with the potential to tap or disrupt US communications with Europe and Asia, received the majority of public attention and have been claimed to be one of the spurs for the creation of NATO Atlantic Command.[23]

B. Space

By contrast with subsea activities, which remain generally invisible, potentially hostile activity in space is more easily documented thanks to its greater visibility to private,

---

19   Andrey Soyustov, "ГУГИ против США: 'скрытая угроза' и невидимый фронт" (GUGI against the USA: the 'hidden threat' and the invisible front), Federalnoye agentstvo novostey, 27 October 2015, https://riafan.ru/455430-gugi-protiv-ssha-skryitaya-ugroza-i-nevidimyiy-front

20   See, for instance, David E. Sanger and Eric Schmitt, "Russian Ships Near Data Cables Are Too Close for US Comfort", *New York Times*, 25 October 2015, https://www.nytimes.com/2015/10/26/world/europe/russian-presence-near-undersea-cables-concerns-us.html

21   Rishi Sunak, "Undersea Cables: Indispensable, Insecure", Policy Exchange, November 2017, https://policyexchange.org.uk/wp-content/uploads/2017/11/Undersea-Cables.pdf

22   Jim Stavridis, "A New Cold War Deep Under the Sea?", *Huffington Post*, 28 October 2015, http://www.huffingtonpost.com/admiral-jim-stavridis-ret/new-cold-war-under-the-sea_b_8402020.html

23   Michael Birnbaum, "Russian Submarines Are Prowling around Vital Undersea Cables. It's Making NATO Nervous", *Washington Post*, 22 December 2017, https://www.washingtonpost.com/world/europe/russian-submarines-are-prowling-around-vital-undersea-cables-its-making-nato-nervous/2017/12/22/d4c1f3da-e5d0-11e7-927a-e72eac1e73b6_story.html; Alexandra Brzozowski, "NATO Seeks Ways of Protecting Undersea Cables from Russian Attacks", Euractiv.com, 23 October 2020, https://www.euractiv.com/section/defence-and-security/news/nato-seeks-ways-of-protecting-undersea-cables-from-russian-attacks/

commercial, and amateur interests involved in or observing space operations. This leads to a preponderance of open source information on threats in space compared to other domains.[24]

In a worst-case scenario, a peer or near-peer adversary could in theory use both land- and space-based anti-satellite (ASAT) weapons systems to launch a mass attack on satellites, targeting the situational awareness of governments and military forces potentially globally, and their ability to communicate, navigate, and target opposing forces – and triggering catastrophic disruption and lasting damage to the space environment. But more discriminate and selective counter-space effects are also possible. Civilian and military communications satellites can be targeted through a wide range of interventions both from ground level and from space itself, including both kinetic and directed-energy attacks.[25] According to General John W "Jay" Raymond, Chief of Space Operations, US Space Force, both Russia and China have "a menu of counter space effects (kinetic, lasers, jammers, cyber)".[26] Iran, North Korea, and India have also developed different techniques to attack or disrupt satellites.[27]

A standard taxonomy of counter-space capabilities includes:

- Co-orbital ASAT;
- Direct Ascent ASAT;
- Electronic Warfare;
- Directed Energy;
- Cyber Attacks.[28]

Co-orbital ASAT capabilities are intended to collide with, damage, or otherwise neutralise their targets. Unusual manoeuvres by Russian space vehicles observed in the vicinity of communications satellites could be practice for attack runs for deploying anti-satellite weapons in order to degrade Western communications at a critical moment,[29] or, in the most charitable explanation, simply an opportunity for close observation and investigation of Western satellites.[30] Russia's Olymp-K or Luch

---

24   See, for example, Brian Weeden and Victoria Samson (editors), "Global Counterspace Capabilities: An Open Source Assessment", Secure World Foundation, April 2020; "Seeking Strategic Advantage: How Geopolitical Competition and Cooperation Are Playing Out in Space", Wilson Center, 6 October 2020, https://www.wilsoncenter.org/event/seeking-strategic-advantage-how-geopolitical-competition-and-cooperation-are-playing-out

25   Leonard David, "China, Russia Advancing Anti-Satellite Technology, US Intelligence Chief Says", Space.com, 18 May 2017, https://www.space.com/36891-space-war-anti-satellite-weapon-development.html

26   Speaking at "Defence Space 2020", 17 November 2020, https://www.airpower.org.uk/defence-space-2020/

27   Todd Harrison et al., "Space Threat Assessment 2020", CSIS, March 2020, https://aerospace.csis.org/wp-content/uploads/2020/03/Harrison_SpaceThreatAssessment20_WEB_FINAL-min.pdf

28   Brian Weeden, "Current and Future Trends in Chinese Counterspace Capabilities", *IFRI Proliferation Papers* 62, November 2020.

29   Patrick Tucker, "Russia Tests a Satellite That Rams Other Satellites, US Says", *Defense One*, 23 July 2020, https://www.defenseone.com/technology/2020/07/russia-tests-satellite-rams-other-satellites-us-says/167154/

30   Brian Weeden, "Dancing in the Dark Redux: Recent Russian Rendezvous and Proximity Operations in Space", *Space Review*, 5 October 2015, http://www.thespacereview.com/article/2839/1

satellite has attracted particular attention by approaching 11 unique Intelsat satellites, four Eutelsat satellites, two SES satellites, and at least nine other satellites operated by Russia, Turkey, Pakistan, the United Kingdom, and the European Space Agency since its launch in September 2014.[31]

By contrast, direct ascent ASAT systems consist of a missile with a kill vehicle launched from land, aircraft, or ship, which collides with the target satellite at high speed and obliterates both objects. Russia has extensively tested weapons of this kind, developed from missile defence systems.[32] And in early 2019, India became the fourth country after the US, China, and Russia to successfully test a ground-launched ASAT missile.[33]

Non-kinetic counter-space capabilities include the use of laser, microwave, and electromagnetic pulse energy against space systems. Anti-satellite EW capabilities can offer interference, denial, and manipulation of radio frequencies operations against satellite and ground support systems.[34] This can also spoof signals from satellites, or simply make it difficult to detect them. Meanwhile, lasers capable of dazzling sensors on satellites could, at greater power, potentially cause physical damage.[35]

And at the juncture of the domains of space and cyber, cyber counter-space operations include capture, disruption, and denial operations against satellite systems through the exploitation of digital vulnerabilities.[36] Unlike electronic attacks, which would prevent satellites communicating, cyber attacks could use the communication channels to deliver corrupted data or malicious commands. Satellite ground stations and their associated communications services would be potential entry points for cyber attacks, while targeting a satellite's command and control system could damage or destroy the satellite, or remove it from orbit.[37] Vulnerabilities to attack have also been found in satellite communications (SATCOM) data links, critically important to military C5ISR

31   Thomas G. Roberts, "Unusual Behavior in GEO: Luch (Olymp-K)", Aerospace Security Project, CSIS, accessed 1 March 2020, https://aerospace.csis.org/data/unusual-behavior-in-geo-olymp-k/
32   "Russia Tests Direct-Ascent Anti-Satellite Missile", US Space Command, 16 December 2020, https://www.spacecom.mil/News/Article-Display/Article/2448334/russia-tests-direct-ascent-anti-satellite-missile/; see also Keir Giles, "Russian Ballistic Missile Defense: Rhetoric And Reality", US Army War College Strategic Studies Institute, June 2015, https://ssi.armywarcollege.edu/russian-ballistic-missile-defense-rhetoric-and-reality/
33   Shaan Shaikh, "India Conducts Successful ASAT Test", Missile Threat, CSIS, 28 March 2019, https://missilethreat.csis.org/india-conducts-successful-asat-test/
34   Todd Harrison et al., "Space Threat Assessment 2018", CSIS, April 2018, https://www.csis.org/analysis/space-threat-assessment-2018
35   Noah Shachtman, "Is This China's Anti-Satellite Laser Weapon Site?" *Wired*, 11 March 2009, https://www.wired.com/2009/11/is-this-chinas-anti-satellite-laser-weapon-site/
36   Rajeswari Pillai Rajagopalan, "Electronic and Cyber Warfare in Outer Space", UNIDIR, May 2019, p. 1–11, https://www.unidir.org/files/publications/pdfs/electronic-and-cyber-warfare-in-outer-space-en-784.pdf; see also Beyza Unal, "Cybersecurity of NATO's Space-based Strategic Assets", Chatham House, July 2019, https://www.chathamhouse.org/2019/07/cybersecurity-natos-space-based-strategic-assets
37   In addition, Russia is believed to have successfully exploited foreign satellites and their unencrypted communications with ground receiver stations as part of a broader cyber campaign. See Sam Jones, "Russian Group Accused of Hacking Satellites", *Financial Times*, September 2015. Available at: https://www.ft.com/content/50b1ff84-571d-11e5-9846-de406ccb37f2

(Command, Control, Communication, Computers, Cyber, Intelligence, Surveillance, and Reconnaissance), transport, industry, and especially aviation technology, where these systems are indispensable.[38] Cyber vulnerabilities in satellite receiving stations also pose secondary risks, as many operational services dependent on data from satellites (for instance, weather services) are distributed via ground station links.[39]

In addition to their effects on civilian communications and other services, targeting of space assets for military effect in conventional conflict is also a substantial risk. US and NATO forces are highly dependent on space-based systems for situational awareness, communication, navigation, and targeting of opposing forces. Degradation or destruction of space assets would put expeditionary forces deploying over long distances at a particular disadvantage relative to the adversary, who would already be present at the edge of the battlespace. Meanwhile, interference with Global Positioning System (GPS) services would negate the effectiveness of GPS-dependent navigation systems and standoff weapons, and dazzling or destruction of surveillance and imaging satellites would prevent observation of the buildup and manoeuvre of adversary forces.

This means that adversaries possessing sufficiently advanced technical capabilities have a strong incentive to target satellites as a key vulnerability.[40] According to Air Chief Marshal Sir Mike Wigston, Chief of Air Staff, RAF, "Future conflict may not start in space, but it may transition quickly to space and it may be won or lost in space".[41] One authoritative assessment of Russian doctrinal and capability developments notes that "Russia considers space as a theater of military operations… Therefore, the emergence of new forms of military operations in near space can be expected".[42] Russia may also view activities in space as a potential component of non-nuclear deterrence, presenting a means of holding high-value adversary targets at risk as an alternative to strategic non-nuclear strike weapons.[43]

C. Land

Denial of access to cyberspace for a targeted region or nation could include physical operations to inflict damage to vital information technology infrastructure on land, such as fibre-optic cables, server farms, terrestrial communication lines, wireless

---

38 Ruben Santamarta, "SATCOM Terminals: Hacking by Air, Sea, and Land", IOActive, 2014, https://www. blackhat.com/docs/us-14/materials/us-14-Santamarta-SATCOM-Terminals-Hacking-By-Air-Sea-And-Land-WP.pdf

39 Mike Gruss, "Report Cites Vulnerability in NOAA's Satellite Ground Stations", *Space News*, August 2014, https://spacenews.com/41685report-cites-vulnerability-in-noaas-satellite-ground-stations/

40 Caroline Houck, "The US Army Knows It's Vulnerable to Space Attack. Here's What They Want to Do About It", *Defense One*, 4 December 2017, http://www.defenseone.com/technology/2017/12/us-army-knows-its-vulnerable-space-attack-heres-what-they-want-do-about-it/144279/

41 Speaking at "Defence Space 2020", 17 November 2020, https://www.airpower.org.uk/defence-space-2020/

42 Timothy Thomas, "Russian Combat Capabilities for 2020: Three Developments to Track", Mitre Corporation, December 2019, https://www.armyupress.army.mil/Portals/7/Legacy-Articles/documents/Thomas-Russian-Combat-Capabilities.pdf

43 Clint Reach, "Review of Strategic Deterrence Book: The Work of Burenok and Pechatnov (2011)", Russia Strategic Initiative, HQ, USEUCOM, 3 December 2020.

communication systems, antennas, telecommunication towers, and associated support infrastructure. By default, contingency planning for civilian facilities of this kind will consider a number of risks such as fire, flood, or intrusion; but resilience to deliberate attack by a well-resourced hostile nation state would entail an entirely different order of security.

Where they exist, single points of failure will be particularly attractive to hostile actors. For several years, internet provision for the entire east of Latvia, including the Latgale region (briefly prominent as a candidate in widely discussed scenarios for a Russian intervention in the Baltic states), reportedly depended on cables under a single bridge across the Daugava river – in the same manner as Crimea's internet access could be controlled by physical intervention at a single point. The aim of this intervention may not be destruction; again, as in Crimea, physical presence inside a trusted facility opens a wide range of possibilities for controlling, selectively interdicting, or manipulating data – or indeed gaining easier remote access to other facilities by appearing to come from inside their security perimeter.

The need for close investigation of potential targets lies behind a sustained effort by Russia to covertly map the United States's telecommunications infrastructure and communications chokepoints,[44] in some instances in suspected coordination with reconnaissance flights carried out by Russian aircraft over the United States under the Open Skies Treaty.[45] In other cases, operations on land spill over from investigating subsea or space targets. Russia has sent covert intelligence officers to Ireland to map precise locations and vulnerabilities where submarine cables linking Europe and America make landfall.[46] Finland in particular has seen media reporting of alarm at the apparently systematic acquisition by Russian interests of land and properties in key locations near strategically important facilities, including "locations related to telecommunication links".[47] The Turku archipelago, in the narrowest stretch of water between southern Finland and Sweden, has been highlighted as a key location where communications cables, energy interconnectors, and strategically important sea lanes are vulnerable.[48] Speculation persists that Russian-owned properties in the

44   Ali Watkins, "Russia Escalates Spy Games after Years of US Neglect", *Politico*, 6 January 2017, https://www.politico.com/story/2017/06/01/russia-spies-espionage-trump-239003

45   Zach Dorfman, "The Secret History of the Russian Consulate in San Francisco", *Foreign Policy*, 14 December 2017, https://foreignpolicy.com/2017/12/14/the-secret-history-of-the-russian-consulate-in-san-francisco-putin-trump-spies-moscow/

46   John Mooney, "Russian Agents Plunge to New Ocean Depths in Ireland to Crack Transatlantic Cables", *Sunday Times*, 16 February 2020, https://www.thetimes.co.uk/past-six-days/2020-02-16/ireland/russian-agents-plunge-to-new-ocean-depths-in-ireland-to-crack-transatlantic-cables-fnqsmgncz

47   Ari Pesonen, "Tietoliikenneyhteyksien katkaiseminen olisi Venäjälle tehokasta sodankäyntiä" (Disconnecting telecommunications would be an effective form of warfare for Russia), *Uusi Suomi*, 27 October 2015, http://aripesonen1.puheenvuoro.uusisuomi.fi/205516-tietoliikenneyhteyksien-katkaiseminen-olisi-venajalle-tehokasta-sodankayntia

48   "Suomen vesiväylät 'motissa' – venäläisfirma osti maat" (Finnish waterways 'in a motti' after Russian company buys land), *Iltalehti*, 19 January 2015, http://www.iltalehti.fi/uutiset/2015011919044524_uu.shtml; "Maakauppoja strategisissa kohteissa", *Iltalehti*, 12 March 2015, http://www.iltalehti.fi/uutiset/2015031119338528_uu.shtml

archipelago raided in a major operation by the Finnish Tax Police, Border Guard, and Defence Forces in late 2018 were intended for use in an interdiction operation as opposed to being simply a non-political money laundering enterprise.[49]

Information interdiction can also be brought about remotely, using Russia's extensive suite of EW capabilities, one of whose key tasks is to "counter the enemy's advantages in the information and telecommunications space".[50] Russia claims that its "Murmansk BN" system deployed on the Kola Peninsula can disrupt communications across northern Europe, with a range of up to 5,000 kilometres.[51] It should not be assumed that the targets for this disruption will be wholly, or even primarily, military: while EW is supposed to achieve the military aims of "delaying timely information support to decision-makers, misguiding them with false information, constructing information blockades, warping databases, and destruction",[52] Russian military thought leaders have also predicted that in the initial period of war, the EW Troops will be tasked with suppressing broadcast and online media, including social media – specifically "blocking radio and television signals, and message traffic in social networks".[53] Russia's capabilities may in fact match its ambition of effecting information interdiction at all levels from individual connected devices such as mobile phones[54] up to national level, affecting broad-scale geographic areas and entities.[55] Both the intent and the capability, and the spillover from military aims to civilian consequences, have been demonstrated by Russia's repeated disruption of GPS navigation provision.[56]

49 Robin Häggblom, "A Dawn Raid in the Archipelago", Corporal Frisk blog, 23 September 2018, https://corporalfrisk.com/tag/airiston-helmi/; see also Joseph Trevithick, "Rumors of Covert Russian Ops Swirl After Finland's Police Raid Bond-Esque Private Island", *The Drive*, 1 November 2018, https://www.thedrive.com/the-war-zone/24616/rumors-of-covert-russian-ops-swirl-after-finlands-police-raid-bond-esque-private-island

50 Yuriy Lastochkin, "Солдаты РЭБ на страже эфира" (EW Troops guarding the airwaves), *Krasnaya Zvezda*, 15 April 2019, http://redstar.ru/wp-content/uploads/2019/04/041-15-04-2019.pdf

51 Jarmo Huhtanen, "Venäjä julkaisi videon, jossa harjoiteltiin häirintä¬järjestelmän käyttöä lähellä Suomen rajaa" (Russia releases video showing training with jamming system near Finnish border), *Helsingin Sanomat*, 13 November 2020, https://www.hs.fi/kotimaa/art-2000007615087.html

52 I. I. Korolyov et al., "Problems in Determining the Methods for Using the Forces and Means of Radio Electronic Warfare as an Arm of the Ground Forces", *Voennaya Mysl'* (Military Thought), No. 9, 2016, pp. 14–17.

53 S. G. Chekinov and S. A. Bogdanov, "Прогнозирование характера и содержания войн будущего: проблемы и суждения" (Forecasting the nature and content of wars of the future: problems and assessments), *Voennaya Mysl'* (Military Thought), No. 10, 2015, pp. 44–45.

54 Kelsey D. Atherton, "Russian Drones Can Jam Cellphones 60 Miles Away", C4ISRNET, November 2018, https://www.c4isrnet.com/newsletters/unmanned-systems/2018/11/16/russian-drones-can-jam-cell-phones-60-miles-away/

55 Martti J. Kari, *Russian Strategic Culture in Cyberspace: Theory of Strategic Culture*, JYU Dissertations 122 (Jyväskylä, Finland: Faculty of Information Technology, University of Jyväskylä, October 2019), 61–63, https://jyx.jyu.fi/bitstream/handle/123456789/65402/978-951-39-7837-2_vaitos_2019_10_11_jyx.pdf

56 Aleksandr Gostev, "'Мишки' на Севере. Был ли российский спецназ на Шпицбергене" ("Little bears" in the north. Was there a Russian Spetsnaz force on Spitsbergen?), Radio Svoboda (Radio Liberty), 2 October 2019, https://www.svoboda.org/a/30195704.html; Kyle Mizokami, "Russia Is Disrupting GPS Signals and It's Spilling into Israel", *Popular Mechanics*, 1 July 2019, https://www.popularmechanics.com/military/weapons/a28250133/russia-gps-signals-israel/

Disruption of GPS has a clear military application in preventing the use of those Western military systems that depend on it for navigation or guidance. But widespread and intensive use of this tactic would also cause severe societal disruption whether within or without an overt conflict due to ubiquitous reliance on positioning, navigation, and timing (PNT) services and the atrophy of skills and services that would replace them.[57] Road movements and every other type of activity that depends on GPS would be hampered; navigation systems without multiple redundancies and fallback systems would be affected, as would millions of embedded systems. Military movements would be impacted even if military navigational systems themselves were resilient; with civilian traffic reliant on GPS, chaos on road networks would be likely. Similarly, in the air, while commercial air traffic would continue to be able to navigate due to redundancy of systems, general aviation with greater reliance on GPS would cause severe ATC and traffic management challenges, for instance by blundering into busy controlled airspace.[58]

D. Online

Finally, adversaries still have the option of destructive effects delivered against information resources remotely through exclusively cyber means. A survey of Chinese cyber activity in the first decade of this century, in addition to intelligence-gathering, identified a range of "activities designed to damage or destroy network elements... as well as infrastructure dependent on those elements, such as communications systems".[59] Continuing concerns over potential hidden payloads in Chinese software, hardware, and firmware drive ongoing debate on the impact on network security of reliance on Chinese providers such as Huawei.[60]

Russia, meanwhile, has developed other means of denying access to the internet for ordinary users, including through exploits such as the VPNFilter malware, capable of permanently disabling home and small office internet connections on demand.[61] Russia's attack on the French TV channel TV5Monde in 2015 included erasing the firmware on nearly all of the network's routers and switches, resulting in blank screens for viewers. A French government investigation concluded that the attackers' primary goal was destruction of the network (and thus its capability to broadcast).[62]

---

[57] As highlighted by Gen. Sir Patrick Sanders of UK Strategic Command, speaking at "Defence Space 2020", 17 November 2020, https://www.airpower.org.uk/defence-space-2020/

[58] See further discussion in Keir Giles, "Missiles Are Not the Only Threat", in *Beyond Bursting Bubbles – Understanding the Full Spectrum of the Russian A2/AD Threat and Identifying Strategies for Counteraction*, FOI, July 2020, https://www.researchgate.net/publication/342643740_Missiles_Are_Not_the_Only_Threat

[59] Desmond Ball, "China's Cyber Warfare Capabilities", Security Challenges, 2011, pp. 81–103, https://www.jstor.org/stable/26461991

[60] "The Security of 5G", House of Commons Defence Committee, Second Report of Session 2019–21 HC 201, https://committees.parliament.uk/publications/2877/documents/27899/default/

[61] Liam Tung, "FBI to All Router Users: Reboot Now to Neuter Russia's VPNFilter Malware", *ZDNet*, 29 May 2018, https://www.zdnet.com/article/fbi-to-all-router-users-reboot-now-to-neuter-russias-vpnfilter-malware/

[62] Matthew J. Schwartz, "French Officials Detail 'Fancy Bear' Hack of TV5Monde", *Bank Info Security*, 12 June 2017, https://www.bankinfosecurity.com/french-officials-detail-fancy-bear-hack-tv5monde-a-9983

This may have formed part of the testing of information warfare capabilities that Russia appeared to be engaged in during the period following Crimea, with the same aim of eliminating competing sources of information – and ensuring that just as in Crimea, governments are unable to communicate with their citizens and populations are denied access to outside information.[63]

3. IMPLICATIONS AND RECOMMENDATIONS

Information warfare in the holistic sense espoused by China and Russia extends far beyond the Western concept of "cyber" activities. As with so many aspects of this challenge, the first and most important task for defenders is recognising the nature and scope of the threat. While many other aspects of information warfare as practiced by adversaries are now much more clearly understood – for instance, the destructive power of disinformation – there has been little public recognition by NATO nations of their adversaries' ambition to deny them use of the internet through physical intervention. Ciaran Martin, formerly founding Chief Executive of the UK's National Cyber Security Centre, classes "adversarial infrastructure destruction" as Level 2 in an ascending five-tier classification of cyber capabilities. But this destruction refers to "persistent engagement" or "counter-cyber" activities delivered through cyberspace and intended specifically to degrade the adversary's cyber capabilities, as opposed to physical activity with broader objectives. Meanwhile, the same classification refers to "kinetic" attacks as Level 4; but here too the discussion is of disruption achieved through cyber rather than physical means. (This classification, interestingly, groups the TV5Monde attack discussed above under "kinetic" impact.)[64]

Once recognition of the specific nature of this challenge is assured, many other countermeasures are familiar from more traditional cybersecurity practice. Given the extent to which the potential targets are in private ownership, defence and security agencies need to foster even closer partnerships with industry in order to access its expertise and secure cooperation at critical moments.[65] Infrastructure owners will be needed to advise on the precise cause of outages in order to inform appropriate responses – to take an example from late November 2019, whether a major outage of e-government services is the result of a cyber attack by a hostile power, or of rats chewing through cables.[66]

---

[63] See extensive discussion of this testing in Keir Giles, "The Next Phase of Russian Information Warfare", NATO Strategic Communications Centre of Excellence, November 2015, https://www.stratcomcoe.org/next-phase-russian-information-warfare-keir-giles

[64] Ciaran Martin, "Cyber-Weapons Are Called Viruses for a Reason: Statecraft and Security in the Digital Age", King's College London, 10 November 2020, https://s26304.pcdn.co/wp-content/uploads/Cyber-weapons-are-called-viruses-for-a-reason-v2-1.pdf

[65] Elisabeth Braw, "National Business Corps to the Rescue", *Foreign Policy*, 23 November 2020, https://foreignpolicy.com/2020/11/23/national-business-corps-to-the-rescue/

[66] "E-services Inaccessible After Rats Chew through Wires", ERR, 21 November 2019, https://news.err.ee/1005241/e-services-inaccessible-after-rats-chew-through-wires

Industry can also assist governments with situational awareness in general. Preparations for many of the attack scenarios described above have protracted timelines. For destructive cyber attacks, preemptive establishment of persistent access to high-value digital and computerised targets can take place long in advance.[67] In space, similarly, "properly positioning an orbital weapon into an appropriate attack position will often take days or weeks".[68] If industry is maintaining an appropriate level of situational awareness, these preparations provide potential opportunities to detect suspicious activity and prepare countermeasures.

Education in awareness of the threat would involve building on current efforts at warning information consumers against disinformation, by informing civilian populations of situations where they may also be receiving apparently trustworthy communications from known sources, including their governments, that are tainted or manipulated as a result of foreign intervention. False messaging on a mass scale, including from previously trusted sources, should be prepared for. Citizens will in many cases find it easier to determine the authenticity of broadcast media than of online information; other NATO nations should consider emulating Latvia, which encourages the public to seek information in time of crisis from television or radio, rather than the internet.[69]

In addition to previous statements by NATO and member states on responses to cyber attacks, declaratory policy should include emphasis that an attack (whether "armed" or not) on critical information and telecommunications assets supporting NATO states would be regarded as a use of force against those states and incur costs accordingly. The ability and will to employ countermeasures against kinetic and non-kinetic attacks should be shown, following the example of French Defence Minister Florence Parly, who in July 2019 promised responses in kind to threats to French space assets.[70]

Meanwhile, the scope for constraint on dangerous activity in or against space through new international agreements seems limited. The rapid development of Russia's capabilities in this field, and its possible advantages over competitors, could account for Russia's position in the United Nations changing over the past decade from proposing arms control treaties in space[71] to opposing a UK initiative on "reducing space threats

---

67 See discussion in "Bearing Witness: Uncovering the Logic behind Russian Military Cyber Operations", Booz Allen Hamilton, 2020.
68 Rebecca Reesman and James R. Wilson, "The Physics of Space War: How Orbital Dynamics Constrains Space-to-Space Engagements", Center for Space Policy and Strategy, October 2020, p. 20, https://aerospace.org/sites/default/files/2020-10/Reesman_PhysicsWarSpace_20201001.pdf
69 Public information video from Latvian State Fire and Rescue Service, 9 May 2019, available at https://twitter.com/ugunsdzeseji/status/1126435222759800833
70 "France to Develop Anti-Satellite Laser Weapons: Minister", *France 24*, 25 July 2019, https://www.france24.com/en/20190725-france-develop-anti-satellite-laser-weapons-minister
71 "Proposed Prevention of an Arms Race in Space (PAROS) Treaty", Nuclear Threat Initiative, 23 April 2020, https://www.nti.org/learn/treaties-and-regimes/proposed-prevention-arms-race-space-paros-treaty/

through norms, rules and principles of responsible behaviours".[72] Furthermore, any meaningful conversation about the future of outer space would require buy-in from all parties involved – including China.[73]

In fact, adversaries willing to target internet infrastructure enjoy a substantial deterrent advantage, as a threat to sow financial or societal chaos through severing undersea cables or jamming GPS might cause a NATO nation to think twice before risking escalation of a confrontation.[74] At first sight, destructive activities against cyberspace might seem self-defeating, since destruction removes access for both the defender and attacker; furthermore, few countries in the world would be immune from the economic repercussions stemming from the impact of such an attack on a major Western power.[75] However, in this respect as in others, Russia has undertaken preparations in the form of efforts to isolate itself from the global internet in time of crisis, with resulting insulation from the blowback effects of any irresponsible activity Moscow might consider undertaking elsewhere.[76]

Instead, more visible deterrence by denial should also form a key part of mitigation strategy for NATO nations. As with all effective means of deterrence, none of the options is cheap or easy; but all are far cheaper and easier than a failure to deter the adversary. Reducing the incentives to target infrastructure could be achieved by demonstrating resilience and redundancy, including publicly developing the capability to operate with a degraded communications environment, which would reduce the perceived benefits of escalation into attacks on civilian systems. Additional measures to improve resilience could include:

- Solutions (albeit expensive and long-term ones) for space vulnerabilities, such as hardening satellites against directed energy attacks and dispensing decoys to confuse direct ascent ASATs.[77]

72 Elena Chernenko, "Звездные войны. Эпизод ООН: Скрытая угроза" (Star Wars: UN episode. The hidden threat), *Kommersant*, 10 November 2020, https://www.kommersant.ru/amp/4565504; "Sending 14 Drafts to General Assembly, First Committee Defeats Motion Questioning Its Competence to Approve One Aimed at Tackling Outer Space Threats", United Nations, 6 November 2020, https://www.un.org/press/en/2020/gadis3658.doc.htm

73 Beyza Unal and Mathieu Boulègue, "Russia's Behaviour Risks Weaponizing Outer Space", Chatham House, 27 July 2020, https://www.chathamhouse.org/2020/07/russias-behaviour-risks-weaponizing-outer-space

74 Katarzyna Zysk, quoted in James Glanz and Thomas Nilsen, "A Deep-Diving Sub. A Deadly Fire. And Russia's Secret Undersea Agenda". *New York Times*, 20 April 2020, https://www.nytimes.com/2020/04/20/world/europe/russian-submarine-fire-losharik.html

75 Louise Matsakis, "What Would Really Happen If Russia Attacked Undersea Internet Cables", *Wired*, 5 January 2018. https://www.wired.com/story/russia-undersea-internet-cables/

76 Juha Kukkola, "Digital Soviet Union", Research Publications No. 40, Finnish National Defence University, 2020.

77 Marcus Weisgerber, "US Air Force Looks For New Ways to Buy, Protect Satellites", *Defense One*, 5 February 2018, http://www.defenseone.com/business/2018/02/us-air-force-looks-new-ways-buy-protect-satellites/145745/

- Ensuring that new communications architectures include redundancies through multiple channels: fibre and cable landlines, mobile networks, and backup and relay stations, including potentially using unmanned aircraft to relay communications.[78]

- Doctrinal and behavioural innovations to reduce reliance on always-on connectivity. Alongside the teaching of media consumer skills in response to disinformation attacks, continuing essential functions by other means when internet access is disrupted or absent should form part of education.

- Preparation and practice by Western governments, and their armed forces, to operate in an environment where communications services normally taken for granted are unavailable. This must include provision and regular exercise of alternative means for distributing public information.

- Explicit inclusion in security and business continuity specifications for critical communications infrastructure of consideration of serious physical attacks – whether carried out by a disaffected conspiracy theorist as in the example that opened this paper, or by an adversary nation state.

Finally, where it is not already the case, both before and during a crisis, civilian internet infrastructure must be accorded the same degree of physical protection as other strategically important assets.

4. CONCLUSION

China, Russia, and other states have developed capabilities which could potentially disrupt or eliminate internet access for NATO states through direct or indirect action against civilian telecommunications infrastructure. Military operations since 2014 demonstrate the availability of telecommunications expertise to Russian special forces in particular, and point to an entirely new integration between cyber, information, and kinetic operations.[79] In effect, the asymmetric information warfare capabilities the Russian Armed Forces aspired to at the beginning of the last decade are now not only available but routinely put to use.[80]

It follows that in time of conflict, declared or undeclared, NATO states may find that access to internet resources may be degraded or entirely absent – including for the

---

[78] Donna Attick, "Robust Communications Relay with Distributed Airborne Reliable Wide-Area Interoperable Network (DARWIN) for Manned-Unmanned Teaming in a Spectrum Denied Environment", Navy SBIR, January 2018, http://www.navysbir.com/n18_1/N181-007.htm

[79] Sydney J. Freedberg, "Army Fights Culture Gap Between Cyber & Ops: 'Dolphin Speak'", *Breaking Defense,* 10 November 2015, http://breakingdefense.com/2015/11/army-fights-culture-gap-between-cyber-ops-dolphin-speak/

[80] Compare Keir Giles, "Information Troops – A Russian Cyber Command?", Proceedings of the Third International Conference on Cyber Conflict, Tallinn, June 2011, https://www.ccdcoe.org/uploads/2018/10/InformationTroopsARussianCyberCommand-Giles.pdf, with discussion of Russian information activities in Syria in Tim Ripley, *Operation Aleppo: Russia's War in Syria* (Telic-Herrick Publications, 2018), p. 192 and passim.

purposes of communicating with their own civilian populations or Armed Forces personnel outside hardened and discrete networks. This applies in equal measure to using any other friendly capabilities which may be compromised by lack of access to the electromagnetic spectrum, including GPS signals. It is essential that conflict and crisis planning by NATO member states recognise this risk and take steps to mitigate it.

In the Same Boat: On Small Satellites, Big Rockets, and Cyber Trust

James Pavur
University of Oxford
Department of Computer Science
Oxford, United Kingdom
james.pavur@cs.ox.ac.uk

Martin Strohmeier
armasuisse
Science + Technology
Thun, Switzerland
martin.strohmeier@armasuisse.ch

Vincent Lenders
armasuisse
Science + Technology
Thun, Switzerland
vincent.lenders@armasuisse.ch

Ivan Martinovic
University of Oxford
Department of Computer Science
Oxford, United Kingdom
ivan.martinovic@cs.ox.ac.uk

Abstract: Launch vehicle "ridesharing" has redefined access to and use of outer space. Today, rockets carry satellites from dozens of countries on shared journeys towards the stars. To ensure that these diverse payloads pose no threat to the overall space mission, safety controls have emerged to protect against mechanical and electrical failure. While these protections were designed to mitigate the risk of probabilistic physical effects, they also have implications for cyber attackers seeking to abuse the trusted status of secondary payloads to harm launch missions.

This paper considers such dynamics through a multidisciplinary lens. It begins by drawing on the perspective of security studies and international relations to characterize what motivates an attacker to target satellite launches. This is combined with a technical analysis which leverages model-based engineering techniques to assess the threat of electronic warfare (EW) and radio frequency interference (RFI) attacks against missile range safety technologies on modern launch vehicles. Through dynamic physical simulation, we demonstrate that even inexpensive nanosatellite platforms have the potential to threaten shared launch vehicles in the hands of motivated cyber adversaries. The paper concludes with a brief discussion

of the implications of these findings for both policymakers and technical researchers interested in cyber-physical threats in orbit.

Keywords: *space, ASAT, threat modeling, cyber-physical, aerospace, critical infrastructure*

1. INTRODUCTION

The emergence of small, low-cost secondary satellite payloads, referred to as CubeSats, has underpinned a revolution in modern space mission design. This has, in turn, reshaped the satellite launch market. Where in the past, rockets carried hardware belonging to a single nation-state or a handful of domestic organizations, today a single launch vehicle may take satellites belonging to dozens of foreign entities on a shared ride to the stars. In this paper, we consider how these trends intersect with the evolving domain of space cyber security.

We take an interdisciplinary approach, starting with an analysis of the global CubeSat launch market and relevant interstate political dynamics. This motivates a novel threat model, leveraging CubeSat payloads as cyber-physical attack vectors against launch operations. We isolate five key CubeSat safety standards which may constrain cyber adversaries but find that most operate under trust assumptions which are vulnerable to malicious circumvention.

Rather than restricting ourselves to high-level strategic threat modeling, we cultivate a baseline intuition for the implications of such malicious safety violations through dynamic physical simulations of a space-to-space radio frequency interference (RFI) attack scenario. The results of these simulations suggest that, even when limited to standard CubeSat components, attackers have wide physical margins within which to cause sustained intentional degradation to safety-critical communications during launch.

This research makes several contributions presenting a novel analysis at the intersection between "launch diplomacy," hardware safety, and cyber security. It represents one of the first attempts to consider the cyber security properties of space launches and, to our knowledge, the first publication to consider space-to-space cyber warfare operations from secondary payloads as a threat vector. Methodologically, this paper demonstrates how policy analysis, model-based engineering methods, and system security techniques can combine to provide cross-domain insights into emerging threats. Finally, the case study, which makes up the latter portion of the paper, serves

as a cautionary example of how safety engineering controls are not necessarily robust to intelligent and strategic adversaries.

2. BACKGROUND

A. CubeSats and Ridesharing

Orbital access is expensive. Even with state-of-the-art technology, single rocket launches can exceed hundreds of millions of dollars (see Table I). To overcome this barrier, satellite owners engage in "ridesharing," purchasing excess capacity on someone else's launch vehicle (LV) for a secondary payload.

TABLE I: EXAMPLE PER-LAUNCH COSTS AND CAPABILITIES OF MODERN LVS

Vehicle	Approx. Launch Cost (USD)	Approx. Mass-to-Orbit (t)
Ariane 5 (ESA)	$150 million [1]	10 (GTO) – 20 (LEO)
Delta IV (NASA)	$300 million [2]	14 (GTO) – 29 (LEO)
Falcon 9 (SpaceX)	$60–100 million [3]	8 (GTO) – 23 (LEO)

Note: GTO = Geosynchronous Transfer Orbit, LEO = Low Earth Orbit

Ridesharing practice has co-evolved with a satellite design template, referred to as CubeSats [4]. CubeSats are small and lightweight, with the smallest size (1 CubeSat Unit or 1U) only 10 cm³ in volume and weighing approximately 1.3 kg. For missions which require large components, multiple 1U cubes can be combined. For example, a 30 × 10 × 10 cm payload weighing around 4 kg would be referred to as a 3U CubeSat.

Compared to traditional satellites, CubeSats are small and cheap, with complete mission costs ranging from tens of thousands to a few million euros [5]. Readymade CubeSat platforms can be purchased online for as little as 25,000 euros, although most missions will require some additional customization [6]. This has made CubeSats the platform of choice for many space start-ups and research missions.

The standard shape and mass of CubeSats allows for easy integration to LVs via standardized deployers, thereby creating a sort of commodity market for global CubeSat launch capacity. The dominant deployer type is the P-Pod (see Figure 1) [4]. A P-Pod is essentially an aluminum box with a door on one end and a spring on the other. When the door's latch is released by the LV flight computer, the spring ejects

up to 3U of CubeSats into space at a velocity of 1–2 m/s. Other deployer types tend to follow similar design principles [7], [8].

FIGURE 1: A 3U CUBESAT (FRONT) AND P-POD (BACK) [9]
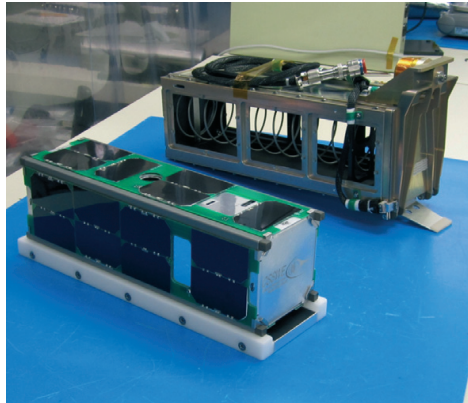
The global rocket launch market to deliver such payloads is consolidated into a handful of major players. Between 2016 and 2019, 90% of the estimated 29 billion US dollars spent on launch services went to one of seven space powers: United States, European Union, China, Russia, Japan, India, and New Zealand [10]. The content of these missions, on the other hand, is highly internationalized. For example, the European Space Agency (ESA) Vega SSMS mission in 2020 delivered a total of 53 satellites to Low Earth Orbit (LEO) [11]. These included platforms for the Thai military, a Russian nuclear physics institute, an Estonian university, and a Facebook subsidiary. In total, 21 customers from 13 countries shared the same journey to the stars.

B. Space Diplomacy and Ridesharing

These multi-state missions occur against a complex geopolitical backdrop. LVs have been longstanding subjects of tension due to their dual-use potential; other than the direction they face, and the logo painted on their side, there is little differentiating an LV from an intercontinental ballistic missile (ICBM). Indeed, both the US and Russia regularly repurpose retired ICBMs for space launches, and responses to North Korea's domestic space program have been inextricably linked to arms control concerns [12]–[14].

The tensions do not stop at the atmosphere's edge. Major military powers rely heavily on space for battlefield communications and operations. As satellites are physically fragile, there is significant fear of attacks on space assets in future conflicts [15]. In particular, prior research has argued that states have strong incentives to engage in cyber attacks due to structural advantages favoring cyber attackers in the space domain [16].

However, there have also been many indications of interstate cooperation. Throughout the Cold War, significant efforts were made by both the US and USSR to cooperate on space launches, giving rise to the Apollo-Soyuz Test Project (ASTP). It has been argued that "track II diplomacy" resulting from interpersonal relationships cultivated during ASTP gave rise to broader diplomatic gains, such as strategic arms control agreements and the demilitarization of Russia's launch sector [17]. In a more modern context, launch collaboration for the International Space Station (ISS) was one of the few aspects of the US-Russia bilateral relationship to survive the diplomatic fallout of Russia's invasion of Crimea in 2014 [18].

Some classical realists treat this sort of cooperation with skepticism. For example, Wang's review of US-EU space cooperation argues that the US used LV ridesharing as a tool to undermine and weaken European rocketry development efforts [19]. Likewise, Chalecki contends that the ASTP was little more than a guise for the US and Soviet military intelligence to spy on each other [20].

In short, satellite ridesharing is as much a geopolitical matter as a technical one. Ridesharing offers direct economic benefits, but it also redirects huge sums of money into foreign aerospace industries and provides political leverage to LV operators that may be unpalatable to some satellite owners.

3. THREAT MODELING

In this context, we can surmise several motivations for cyber attackers to target launches. A launch failure could prevent or delay the deployment of key space assets. Moreover, commercial actors may see a benefit in harming the reputation of key competitors. For example, this was briefly investigated as a possible cause of a 2016 SpaceX rocket explosion [21]. The prestige and economic importance of space programs may also make them attractive targets for hostile states – as Russian officials suggested following a string of rocket failures in the early 2010s [22], [23].

For this paper, we focus on threats involving the compromise of an inexpensive CubeSat secondary payload. We propose four reasons CubeSats may represent attractive targets.

1) Heavy use of commercial-off-the-shelf (COTS) components allows attackers to develop exploits on representative hardware or software. This contrasts with larger platforms which tend to rely on bespoke components.

2) The COTS supply chain can be compromised; for example, through a backdoor in an open-source software library or the online sale of a malicious sensor. The high number of CubeSats per LV increases the odds of a backdoored product ending up attached to an LV of interest to an attacker.

3) While large satellites and LVs are typically built by nation-states and defense contractors, CubeSats frequently come from start-ups or universities. These organizations are comparatively permeable to digital compromise, insider threats, sabotage, and social engineering.

4) CubeSats are inexpensive. Combined with the pseudo-commodity market for CubeSat launch slots, a proxy corporation or state-sponsored university could afford many attempts at building and launching a CubeSat with malicious flight software that abuses trusted/approved COTS components to cause harm.

To date, little prior technical research exists on CubeSat cyber security in large part due to their low capabilities and small size. To quote one CubeSat developer: "What's the worst that could happen? […] With no propulsion and no pointing control, it's very likely that you couldn't do anything other than turn the camera off" [24]. CubeSat manufacturers have lobbied against cyber-security standards, contending that they pose "an excessive and unnecessary burden, and a major potential mission-reliability risk" [24]. The effect of this mentality is that CubeSats tend to forgo security to meet aggressive cost and schedule requirements. Additionally, in a high-level review of CubeSat security practices, Ingols and Skowyra note that CubeSat developers will often "conflate reliability engineering with security engineering" [25, p. 11].

This is an important point, because while security risks are frequently dismissed, attackers may still struggle to cause meaningful harm after successfully compromising a CubeSat. CubeSats represent many organizations' first space mission and, as a result, often fail. Roughly 50% of CubeSats suffer "infant mortality," failing within six months, and one in five are "dead on arrival," never making contact with Earth at all [26], [27]. Launch providers are thus keenly aware of the risks of strapping unreliable novice hardware, however small, to a cylinder full of rocket fuel. This has given rise to extensive controls designed to limit the mechanical and electrical risk a CubeSat can pose to the LV. In Section 4, we will consider these safety controls and their implications for an intelligent cyber adversary.

4. ADVERSARIAL ANALYSIS OF LAUNCH SAFETY CONTROLS

CubeSat safety requirements can vary substantially and revolve around a series of mission-specific Interface Control Documents (ICDs) provided by the mission integrator. These requirements are complex and certification is non-trivial; NASA recommends allowing 18 months for certification and licensing [28]. In this paper, we focus on two dominant standard documents (among myriad) for CubeSat missions: the *CubeSat Design Specification, REV 13* (CDS) and the *Air Force Space Command Manual 91-710, Volume 3* (AFSPCMAN) [29], [30].

A. CubeSat Design Specification (CDS)

The CDS focuses mostly on the physical properties which may impact a CubeSat's ability to deploy smoothly from a P-Pod. Beyond this, it imposes three broad categories of controls which appear to constrain cyber adversaries.

First, CDS requires deployment switches, small pins on CubeSat rails which are depressed while the CubeSat sits in its P-Pod [29, Sec. 3.3]. These electrically isolate the CubeSat's flight computer from power during launch to prevent a CubeSat from deploying hardware in the P-Pod. They also prevent attackers from launching software-based attacks prior to deployment. Second, CDS prohibits CubeSats from transmitting radio signals until 45 minutes have elapsed from deployment, although the CubeSat may boot up and perform other tasks in that time [29, Sec. 3.4]. This mitigates the risk of both unintentional and malicious radio frequency interference (RFI). Third, CDS typically limits stored chemical energy to 100 watt-hours [29, Sec. 3.1]. This limits the available power for direct physical harm – such as deliberate overheating of key components.

These controls are normally verified using three mechanisms [28]. Battery characteristics are outlined in a battery report, which details specific part numbers and modifications. Radio and electrical interrupts are summarized in an electrical report containing circuit diagrams. Finally, inhibits are verified during a Day in the Life (DITL) test. In a DITL, the CubeSat runs through a simulated separation and a timer is used to verify that no premature transmissions take place. The DITL is typically conducted by the CubeSat developer in their own lab [31], [32].

B. Air Force Space Command Manual 91-710 (AFSPCMAN)

AFSPCMAN consists of more than 200 pages of requirements for launch operations, the primary purpose of which is range safety. The objective of range safety is to protect individuals, vehicles, and structures from harm and ensure that rockets adhere to intended trajectories. Range safety violations can result in the initiation of a self-

destruction system known as a Flight Termination System (FTS), which is designed to ensure that a launch vehicle combusts fully prior to colliding with the Earth's surface.

The primary AFSPCMAN burden for CubeSat developers is the provision of a Missile System Prelaunch Safety Package (MSPSP), prepared by the CubeSat developer [30, p. 214]. It consists of a detailed description, including schematics and functional diagrams, of the payload and relevant hazards.

The most obviously applicable portion of AFSPCMAN to cyber security is the portion on Computer Systems and Software [30, p. 200]. Software security requirements are derived from Software Criticality Indexes (SwCIs) specified in MIL-STD-882E [33]. A synthesis of these requirements can be found in Table II. In most cases, CubeSat software falls in the range of SwCI 4-5, with DITL testing meeting validation burdens. The only additional software safety hurdle is likely a descriptive overview of computing hardware components and software logic [34].

TABLE II: OVERVIEW OF AFSPCMAN SOFTWARE SAFETY STANDARDS

Software Control Category	Severity Level of Safety Failure			
	Catastrophic *(e.g., loss of life, > $10M damages)*	**Critical** *(e.g., hospitalization of 3+ personnel, > $1M damages)*	**Marginal** *(e.g., injury causing lost workdays, > $100K damages)*	**Negligible** *(e.g., minor injury, < $100K damages)*
Autonomous	SwCI 1 (Code Review)	SwCI 1 (Code Review)	SwCI 3 (Architecture Review)	SwCI 4 (Safety-Specific Testing)
Semi-Autonomous	SwCI 1 (Code Review)	SwCI 2 (Design Review)	SwCI 3 (Architecture Review)	SwCI 4 (Safety-Specific Testing)
Redundant Fault Tolerant	SwCI 2 (Design Review)	SwCI 3 (Architecture Review)	SwCI 4 (Safety-Specific Testing)	SwCI 4 (Safety-Specific Testing)
Influential / Informational	SwCI 3 (Architecture Review)	SwCI 4 (Safety-Specific Testing)	SwCI 4 (Safety-Specific Testing)	SwCI 4 (Safety-Specific Testing)
No Safety Impact	SwCI 5 (No Analysis)	SwCI 5 (No Analysis)	SwCI 5 (No Analysis)	SwCI 5 (No Analysis)

Note: The controls in this table are synthesized from multiple tables in *MIL-STD-882E* and controls in *AFSPCMAN 91-703v3* [30], [33]. All controls which apply to lower severity Software Criticality Indexes (SwCI) apply to high severity indexes cumulatively.

Beyond software safety, the MSPSP also imposes requirements to mitigate the risk of electromagnetic interference. A CubeSat developer typically must provide a transmitter survey, which lists all radio transmitters and their fundamental characteristics. This includes an outline of frequency ranges, bandwidth, and deployed and maximum power delivery to a given antenna [28].

Range safety may require verification of emission characteristics through measurements conducted by an approved representative [30, p. 43]. However, in practice, CubeSat missions may be able to avoid the costs and scrutiny of such assessment through the use of RF power inhibits [34]. If frequency analysis is required, the main purpose is to ensure that payload emissions do not broadcast on key frequencies outlined in the LV's specification. These frequencies are often listed in public documentation and typically consist of telemetry and FTS modules [35], [36].

C. Adversarial Analysis

Initially, these controls appear to severely constrain an attacker's capabilities. However, their implementation assumes an informed and benign CubeSat developer who shares the launch integrator's desire for a successful mission.

Under our adversarial model, this shared priority does not exist. CubeSat developers may be unaware of or complicit in efforts to circumvent controls. As large parts of the certification process are self-reported, violating controls is often little more than a matter of ticking an incorrect box or writing down inaccurate numbers on a form. Attackers can strategically evade only a small subset of the hundreds of standards, maximizing potential harm while minimizing detectability.

For example, Ingols and Skowyra note that CubeSats spend the months between completion and launch being passed around different storage facilities and may be subject to post-certification tampering via social engineering vectors [25]. A sophisticated attacker may make minor software modifications to devices during this time with little risk of detection. Even more severely, if the CubeSat developer misrepresents DITL results or electrical diagrams, there is no clear mechanism for detecting this; CubeSats are too fragile to disassemble for manual inspection.

These deceptions may be intentional, or they may come from further up the supply chain. A malicious COTS vendor, for instance, might provide a telemetry module which they purport to broadcast in certain launch compatible frequencies when, in fact, a hidden backdoor enables transmissions in prohibited bands or power levels.

In Table III, we present a demonstrative analysis of five controls from CDS and AFSPCMAN under adversarial conditions. These controls were selected as likely

targets for attackers seeking to cause cyber kinetic harm within the constraints of CubeSat hardware. For each, we note the source of verification authority and deception exposure to both insiders and outsiders.

TABLE III: ADVERSARIAL CIRCUMVENTION ANALYSIS FOR SELECTED SAFETY CONTROLS

Safety Control	Primary Reference	Responsible for Verification	Likely Vulnerability to Malicious Outsider	Likely Vulnerability to Malicious Insider
Deployment switches prevent power-on in deployer	CDS 3.3	CubeSat Developer (DITL, Electrical Diagrams)	**Low** *CubeSat developer would likely detect unauthorized power draw during DITL.*	**High** *CubeSat developer could forge documentation and DITL results.*
Software timers prevent RF transmission for 45 minutes	CDS 3.4	CubeSat Developer (DITL)	**Moderate to High** *Otherwise trivial modifications to code may necessitate special effort to evade DITL detection.*	**High** *CubeSat developer could forge DITL results or program DITL behavior to differ from launch.*
Battery power limitation	CDS 3.1	CubeSat Developer (Battery Report, MSPSP)	**Low** *Malicious vendor could misrepresent battery specs but targeting is logistically complex.*	**Low to Moderate** *Weight and physical properties act as limits on plausible extent of deception.*
Software Safety Guidance	AFSPCMAN A2.2.4.14	CubeSat Developer (MSPSP)	**High** *Software, especially third-party libraries, is unlikely to be audited beyond cursory summary in MSPSP.*	**High** *CubeSat developer will likely only need to provide easily falsified summary information on software operations and design.*
RF Emission Compatibility	AFSPCMAN A2.2.4.10.2, Launch Vehicle User's Guide	CubeSat Developer (MSPSP) Range Safety (EMF testing)	**Low to Moderate** *Malicious vendor could backdoor telemetry hardware. If a software defined radio (SDR) is used, attacker may modify configuration using code.*	**Moderate to High** *In the absence of independent EMF testing, CubeSat developer can lie. Otherwise, they may modify code to change behavior under test conditions.*

This analysis suggests that many of the controls which help ensure safety during the CubeSat integration process are not robust to an intelligent adversary. For example, requiring triple-redundant radio inhibits (CDS 3.4) dramatically reduces the risk from equipment failure. However, there is little difference from the perspective of a malicious CubeSat developer lying once in their electrical report versus lying thrice. Even absent insider access, the lack of software and supply-chain auditing processes provides ample opportunity for cyber attackers to circumvent key safety requirements.

5. THREAT SIMULATION AND EVALUATION

Given the common perceptions that a CubeSat's low capabilities mean that, even in the event of full compromise, it cannot pose a physical threat, it is worth considering the specific technical implications of malicious safety control violations. To do this, we will replicate a hypothetical attack scenario through dynamic physical simulation. The intent is not to completely model the behavior of LVs and satellites but rather to evaluate the general plausibility of harm from compromised CubeSat hardware during launch.

Our hypothetical threat scenario focuses on GPS interference attacks for three reasons. First, RFI attacks are intuitively bolstered by physical proximity – one of the main boons from compromising a secondary payload. Second, what limited public information is available on LV FTS hardware makes it clear that GPS is a key data source [37]. Finally, due to US commercial radio licensing regulations, there is a relative abundance of technical data regarding representative radio hardware, helping to better ground our simulations [38].

A. Scenario Overview

The compromised CubeSat in our simulation is summarized in Table IV. It consists of a notional 3U commercial payload, weighing 4 kg and scheduled for launch on a SpaceX Falcon Heavy. The mission sequence is loosely modeled on that of the STP-2 launch. STP-2 is selected as an example of a mission which deployed CubeSats *en route* to delivery of the primary payload. This emerging practice offers commercial and logistical benefits, but also raises the risks from compromise as CubeSats are deployed while the primary payload and substantial fuel quantities remain in the LV.

Our attacker is derived from the insider model in the rightmost column of Table III. It is a malicious state-sponsored business that has built a CubeSat with the express purpose of circumventing key safety controls. To reduce scrutiny, the attacker is restricted to standard CubeSat components. There are two relevant hardware modules used in the attack, both belonging to the CubeSat's Telemetry, Tracking and Control (TT&C) subsystem.

First, the CubeSat leverages a software defined radio (SDR) transceiver. Specifically, we have modeled our simulation around the 1U μSDR-C from Space Micro [39]. An SDR permits the attacker to dynamically alter radio transmission parameters, including carrier frequencies, using undisclosed software logic. SDRs are commonly used in CubeSats and the presence of an on-board SDR alone would be unlikely to arouse suspicion. Additionally, the attacker has selected an antenna with undisclosed operability in the 1.1–1.6 GHz range as well as the allocated TT&C band. This can

be achieved with a customized deployable antenna, a multi-band module, or an ultra-wideband offering [40]–[42]. This frequency range is selected due to its potential to cause interference with GPS reception.

TABLE IV: ATTACKER CUBESAT CHARACTERISTICS AND OBJECTIVES

Size & Weight	Relevant RF Range	Attacker RF Tx Power	Attacker Objective	Targeted Frequencies	Effective Power at LV
3U, 4 kg	1.1–1.6 GHz	1–10 W	Interfere with L-Band GNSS reception on launch vehicle	1575.42 MHz (GPS Rx)	Approx. -120 dBm (Rx) [43]

The attacker has also inserted malicious programming logic with the intention of circumventing two safety controls from Table III. First, the attacker will begin RF transmission immediately after separation from the P-Pod, violating the 45-minute silence mandate. Second, the attacker will transmit on frequencies prohibited by AFSPCMAN A2.2.4.10.2 and the Falcon User's Guide [35]. To evade detection during lab certification and DITL tests, this malicious logic will check the measurements of on-board sensors (e.g., a thermometer) and only trigger the attack when conditions match LEO.

The attacker's goal is to introduce RFI of sufficient magnitude to trigger a range safety incident on the LV. For example, if positional telemetry data is unavailable or indicates a rocket has strayed from its intended trajectory, this can lead to a mission abort.

This is particularly relevant for the Falcon Heavy, as it is one of the first LVs to include a fully autonomous flight termination system (AFTS) [35, p. 8]. This AFTS can automatically self-destruct the launch vehicle without human approval if sensors show a deviation from approved mission parameters. Although the precise AFTS specifications are, unsurprisingly, restricted, NASA documents confirm GPS observations as a key decision metric for termination [37].

B. Experimental Design and Assumptions

The primary purpose of these simulations is to determine the plausible limits of CubeSat hardware to emit an RF signal which causes sustained degradation to GPS reception. In practice, many relevant dynamics are mission-dependent, such as antenna directionality, GPS satellite locations, and precise launch trajectories. Here, we focus on a "worst case" scenario based on typical GPS signal characteristics, idealized isotropic antennas, and the assumption of equivalent receiver gain across legitimate and illegitimate transmission sources.

1) Model Parameters

According to the Falcon User's Guide, the launch vehicle contains GPS receivers which operate in the L1 signal band (1574.2 MHz) [35]. To determine the necessary jammer characteristics to cause disruption to these signals, we must approximate the strength of legitimate signals at the receiver. The GPS specification only provides information regarding the Earth's surface, but we can derive a more accurate value for LEO. One method for doing so is presented in [43], suggesting an approximate received power of around -120 dBm in dynamic simulation. This is fairly close to the value predicted by a simple Free Space Path Loss (FSPL) model on the basis of the public GPS L1 link budget – with minor modification to account for LEO conditions (see Equations 1–3) [44].

Letting:

$$FSPL(dB) = -10 \times \log_{10}\left[\left(\frac{4\pi d}{\lambda}\right)^2\right] \tag{1}$$

and:

$$P_{rcvr}(dBm) = EIRP_{TX} + FSPL - 30 \tag{2}$$

where:
$d = distance\ from\ transmitter \sim 19{,}000$ meters (varies depending on orbit and time)
$\lambda = wavelength \sim 0.19$ meters
$EIRP = effective\ isotropic\ radiated\ power \sim 26.50$ dBW

$$P_{rcvr}(dBm) = 26.50 \pm 10 \times \log_{10}\left[\left(\frac{4\pi 19000}{0.19}\right)^2\right] - 30 = -125.48\ dBm \tag{3}$$

We can supplement this theoretical analysis with experimental data from the US Department of Transportation (DOT) [45]. Through anionic chamber measurements evaluating the threat of interference from cellular LTE towers (at 1530 MHz) on LEO GPS reception, DOT calculated a receiver threshold of -73 dBm for near-band interference on two NASA platforms [45, p. 110]. As our attacker can jam directly in the L1 band, rather than the adjacent LTE frequencies, we can reasonably assume equivalent or greater interference at this threshold.

2) Simulation Process

Our physical simulation consists of two sub-components – an astrodynamics model for CubeSat separation and an RF interference model. In the astrodynamics model, we replicate the separation of a CubeSat from a P-Pod deployer into LEO. This is implemented in FreeFlyer, a commercial space mission planning tool [46]. The CubeSat ejects from the launch vehicle through a contra-velocity maneuver at 2 m/s as is typical for a 4 kg CubeSat [47]. The CubeSat and launch vehicle are propagated

for a two-hour period following separation, and a separation vector is calculated between the two objects at regular one-minute intervals.

These separation vectors are then leveraged in RF interference simulations. We replicate RF dynamics using MATLAB's Antenna Toolbox, a commercial communications system simulation and development toolkit [48]. Two transmitters are modeled: an L1 GPS transmitter based on the aforementioned P_{rcvr} characteristics and a CubeSat jammer with varying EIRPs from 1–10W. A GPS receiver is replicated on board the rocket. Antenna positions are derived based on the separation vectors calculated in the astrodynamics model and used to compute signal-to-interference-plus-noise ratios (SINR) and $P_{jammer\ at\ rcvr}$(dBm) at regular one-minute intervals.

Under benign conditions ($P_{jammer\ at\ rcvr}$(dBm) = 0), our model computes: $P_{gps\ at\ rcvr}(dBm)$= -125.48, and $SINR(dB)$= -21.41. These values align with our analysis in Section 5.B.1 and prior work, suggesting reasonable fidelity [43], [49].

C. Results and Evaluation

Figure 2 summarizes the output of our astrodynamics model. Note that the separation vector of magnitude does not increase linearly. This is a result of the relative orbital motion of the CubeSat and LV, both of which are in LEO at time of deployment. In our threat model, the attacker does not adhere to the 45-minute radio silence window mandated by the CDS. This means that it can jam immediately after separation and at close proximity to the LV.

FIGURE 2: CUBESAT SEPARATION FROM LV OVER TIME
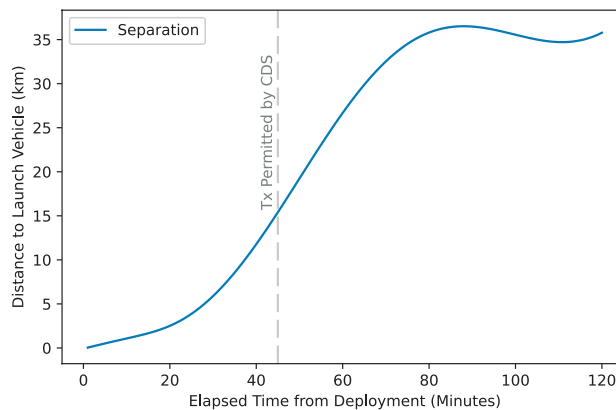
Incorporating these results into our interference model shows that the attacker is capable of degrading GPS signal quality (see Figure 3). As expected, the attack is most effective at higher power levels and during the first few minutes following separation.

Using the aforementioned DOT near-band threshold of -73 dBm gives a conservative estimate of 20–40 minutes of disruption depending on amplifier power (see Figure 4).

FIGURE 3: SINR AT LV RECEIVER DURING ATTACK
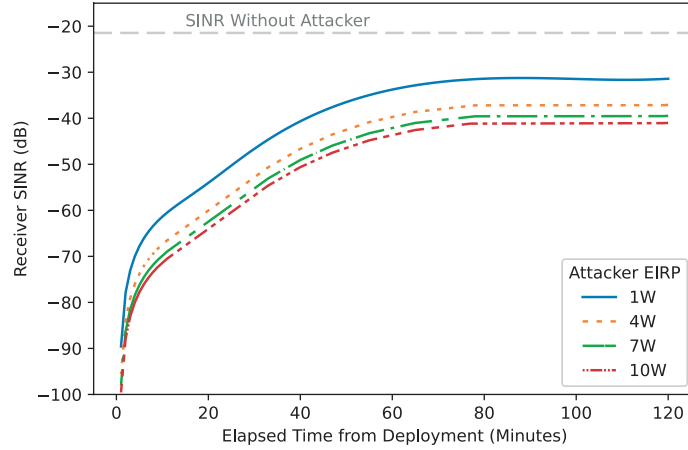
FIGURE 4: JAMMER POWER AT LV RECEIVER DURING ATTACK
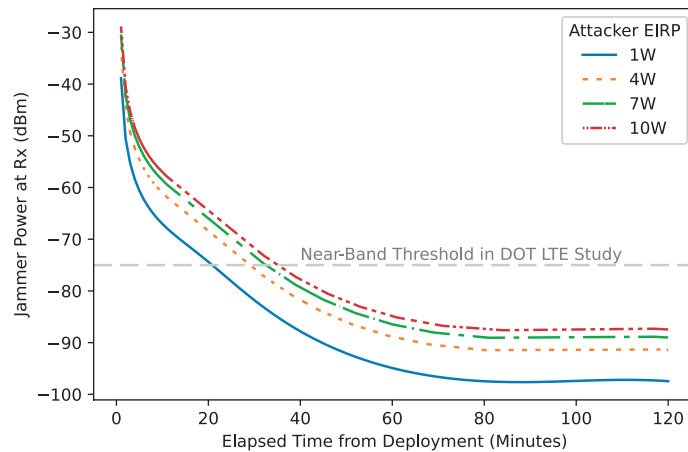
To further validate these bounds, we can convert the SINR to Carrier-to-Noise-plus-Interference Density ratio C/N_{O+I}, assuming a typical front-end bandwidth (BW) of 4e6 Hz and applying the conversion method presented in [49] and in Equation 4:

$$C/N_{O+I}(dB{-}Hz) = SINR + 10 * \log_{10}(BW) \qquad (4)$$

Standard GPS L1 receivers function at C/N_Os between 35 and 55 dB-Hz, with complete loss of signal acquisition below 28 dB-Hz – although this can vary depending on specific hardware conditions [50]. This suggests that our attacker can have a severe impact on GPS quality, keeping C/N_{O+I} below 28 dB-Hz for upwards of 45 minutes at low EIRPs and throughout the simulated period at higher EIRPs (SINR ≤ -38 dB). It

may be prudent to assume that GPS receivers on LVs have access to the wider 20.46 MHz P(Y) frequencies restricted for military use. If this were the case, an attacker would be weaker, but could still expect 30–60 minutes of successful disruption (SINR ≤ -45 dB).

In short, these results suggest it is physically plausible for COTS CubeSat hardware to introduce meaningful disruptions to LV GPS reception on the scale of tens of minutes to several hours depending on mission hardware. While operationalizing such an attack would take significant effort, the low cost and accessibility of CubeSat hardware and launch capacity make it well within the means of state-sponsored attackers. Moreover, the reputational risk of attack failure or attribution is limited as key forensic evidence of the attack would be trapped 1,000 km in the sky.

D. Mitigations and Future Work

The scenario considered here is but one of many possible manifestations of our threat model. The underlying vulnerability proposed here has less to do with GPS reception than with the implicit trust dynamics in secondary payload integration. One promising avenue for future work might thus be to build on this adversarial analysis to identify other technical attack vectors of interest (e.g., premature hardware deployment to jam P-Pod deployers).

Our own RFI scenario also leaves room for future work. Due to limited public information, we could not account for the specific AFTS design. AFTS systems may already have a variety of undocumented defenses, such as leveraging multi-constellation GNSS data, elevating the importance of accelerometer readings in the case of GNSS anomalies, or employing various jamming resistance techniques [51]. To the extent that such mitigations are not implemented, they also represent feasible technical steps towards mitigating the attacks proposed here.

At a high level, our research suggests ample opportunity for future work in adjusting trust models around CubeSat integration policies. This is complex, as CubeSats are built under aggressive timeline and budgetary constraints. However, certain properties – such as the validity of hardware interrupts, operational frequencies of RF hardware, or behavior during DITL testing – may be of sufficient importance to merit the added cost of third-party validation. Launch operators may consider offering expedited certification routes for certain pre-approved COTS components, such as antennas which lack capabilities in sensitive frequencies, to reduce compliance costs. Similarly, they may consider allowing developers to gain trust over time, easing the pathways to large-scale CubeSat deployments while still mitigating the risks from naïve or fraudulent first-time developers.

In short, a comprehensive review of the existing integration certification process from an adversarial perspective is beyond the scope of this paper but represents an intuitive next step for launch operators and regulators concerned about potential harm from compromised or malicious third-party payloads.

6. CONCLUSION

In this paper, we have presented the case that strong political and strategic motivations exist for attacks targeting space launch missions. Moreover, we present, to our knowledge, the first cyber-physical threat model targeting LVs through a secondary payload.

While existing CubeSat safety standards employed in the integration and certification process initially appear to constrain cyber adversaries, we find that unverified trust assumptions underpin the real-world practice of this safety qualification process. When considered in the context of a sufficiently motivated malicious cyber adversary, many safety protections appear trivially circumventable.

The implications of this are evaluated experimentally through physical simulations of a novel space-to-space radio interference attack scenario targeting a modern LV. Our results demonstrate that inexpensive CubeSat hardware has sufficient physical capabilities to potentially threaten the reliability of key safety metrics during launch. We further considered how future work might identify related attacks against other launch systems and isolated steps towards mitigating both this specific attack and others of this nature.

For hundreds of satellite operators, transnational launch collaboration has brought space closer than it has ever been. It offers access for start-ups, universities, and states who would otherwise be unable to reach orbit. Moreover, it fosters key links for communication and diplomacy between scientists and engineers in otherwise deeply sensitive domains. However, trust is a keystone component of sustained cooperation. Ensuring security against both cyber and physical risks will be critical to reaping sustained benefits from globalized launch services.

REFERENCES

[1] E. Howell. "Ariane 5 Rocket Lofts 2 Satellites on Milestone 100th Launch." Space.com. https://www. space.com/41936-ariane-5-rocket-aces-100th-launch.html (accessed Sep. 22, 2020).
[2] E. Howell. "Delta IV Heavy: Powerful Launch Vehicle." Space.com. https://www.space.com/40360-delta-iv-heavy.html (accessed Sep. 22, 2020).

[3]   M. Sheetz. "Elon Musk touts low cost to insure SpaceX rockets as edge over competitors." CNBC.com. https://www.cnbc.com/2020/04/16/elon-musk-spacex-falcon-9-rocket-over-a-million-dollars-less-to-insure.html (accessed Sep. 22, 2020).

[4]   J. Puig-Suari, C. Turner, and W. Ahlgren, "Development of the standard CubeSat deployer and a CubeSat class PicoSatellite," *2001 IEEE Aerospace Con. Proc. (Cat. No.01TH8542)*, vol. 1, pp. 1/347–1/353, Mar. 2001, doi: 10.1109/AERO.2001.931726.

[5]   E. Kulu. "Nanosats Database." Nanosats Database. [Online]. Available: https://www.nanosats.eu/tables.html (accessed Sep. 24, 2020).

[6]   EnduroSat. "1U CubeSat Platform." EnduroSat.com. https://www.endurosat.com/cubesat-store/all-cubesat-modules/1u-cubesat-platform/ (accessed Sep. 24, 2020).

[7]   Nanoracks. "ISS Deployment." Nanoracks.com. https://nanoracks.com/products/iss-deployment/ (accessed Sep. 24, 2020).

[8]   Innovative Solutions in Space. "ISIS ISIPOD 3-Unit CubeSat deployer." CubeSatShop.com. https://www.cubesatshop.com/product/3-unit-cubesat-deployer/ (accessed Sep. 24, 2020).

[9]   CSSWE, *English: The CSSWE CubeSat and PPOD just prior to integration*, 2012. [Online Image]. Available: https://commons.wikimedia.org/wiki/File:CSSWE_CubeSat_and_PPOD_prior_to_integration.png. License: CC-BY-SA-3.0.

[10]   C. C. Helms, "A Survey of Launch Services 2016–2020," in *AIAA Propulsion and Energy 2020 Forum*, AIAA, 2020.

[11]   eoPortal. "Vega PoC flight for SSMS." 2020. https://directory.eoportal.org/web/eoportal/satellite-missions/v-w-x-y-z/vega-ssms (accessed Dec. 2, 2020).

[12]   Northrop Grumman. "Minotaur Rocket." NorthropGrumman.com. https://www.northropgrumman.com/space/minotaur-rocket (accessed Dec. 2, 2020).

[13]   W. Graham. "Russia's Rokot vehicle successfully launches Geo-IK-2 satellite." NASASpaceFlight.com. https://www.nasaspaceflight.com/2019/08/russias-rokot-geo-ik-2-satellite/ (accessed Dec. 3, 2020).

[14]   P. Olbrich and D. Shim, "Symbolic practices of legitimation: exploring domestic motives of North Korea's space program," *Int. Relat. Asia Pac.*, vol. 19, no. 1, pp. 33–61, Jan. 2019, doi: 10.1093/irap/lcx004.

[15]   Defense Intelligence Agency, "Challenges to Security in Space," 2019. [Online]. Available: https://www.dia.mil/Portals/27/Documents/News/Military%20Power%20Publications/Space_Threat_V14_020119_sm.pdf

[16]   J. Pavur and I. Martinovic, "The Cyber-ASAT: On the Impact of Cyber Weapons in Outer Space," in *2019 11th Int. Conf. on Cyber Conflict (CyCon)*, 2019, vol. 900, pp. 1–18.

[17]   J. C. Mauduit, "Collaboration around the International Space Station: science for diplomacy and its implication for US-Russia and China relations," *Proc. 7th Ann. SAIS Asia Conf. (SAIS 2018)*, Feb. 17, 2017. [Online]. Available: https://swfound.org/media/205798/sais-conference-jcmauduit-paper.pdf

[18]   M. Byers, "Cold, dark, and dangerous: international cooperation in the arctic and space," *Polar Record,* vol. 55, no. 1, pp. 32–47, Jan. 2019, doi: 10.1017/S0032247419000160.

[19]   S.-C. Wang, "The Making of New 'Space': Cases of Transatlantic Astropolitics," *Geopolitics*, vol. 14, no. 3, pp. 433–461, Aug. 2009, doi: 10.1080/14650040802693820.

[20]   E. L. Chalecki, "Knowledge in Sheep's Clothing: How Science Informs American Diplomacy," *Diplomacy & Statecraft*, vol. 19, no. 1, pp. 1–19, Mar. 2008, doi: 10.1080/09592290801913676.

[21]   C. Davenport. "Implication of sabotage adds intrigue to SpaceX investigation." *The Washington Post*. Sep. 30, 2016. https://www.washingtonpost.com/business/economy/implication-of-sabotage-adds-intrigue-to-spacex-investigation/2016/09/30/5bb60514-874c-11e6-a3ef-f35afb41797f_story.html (accessed Dec. 11, 2020).

[22]   RT International. "Sabotage considered in Proton rocket crash – investigator." RT.com. May 29, 2014. https://www.rt.com/news/162228-proton-rocket-failure-sabotage/ (accessed Dec. 11, 2020).

[23]   F. Weir, "Russia hints foreign sabotage may be behind space program troubles," *Christian Science Monitor*, Jan. 10, 2012. https://www.csmonitor.com/World/Global-News/2012/0110/Russia-hints-foreign-sabotage-may-be-behind-space-program-troubles/ (accessed Dec. 11, 2020).

[24]   D. Werner. "Small satellite sector grapples with cybersecurity requirements, cost." SpaceNews.com. August 8, 2018. https://spacenews.com/small-satellite-sector-grapples-with-cybersecurity-requirements-cost/ (accessed Sep. 21, 2020).

[25]   K. W. Ingols and R. W. Skowyra, "Guidelines for Secure Small Satellite Design and Implementation: FY18 Cyber Security Line-Supported Program," MIT Lincoln Laboratory Lexington United States, Feb. 2019. Accessed: Sep. 24, 2020. [Online]. Available: https://apps.dtic.mil/sti/citations/AD1099003

[26]   M. Langer and J. Bouwmeester, "Reliability of CubeSats – Statistical Data, Developers' Beliefs and the Way Forward," *Proc. 30th Ann. AIAA/USU Conf. Small Satell.,* 2016, [Online]. Available: https://repository.tudelft.nl/islandora/object/uuid%3A4c6668ff-c994-467f-a6de-6518f209962e

[27] M. Swartwout, "You say 'Picosat', I say 'CubeSat': Developing a better taxonomy for secondary spacecraft," in *2018 IEEE Aerosp. Conf.*, pp. 1–17, Mar. 2018, doi: 10.1109/AERO.2018.8396755.

[28] NASA, *CubeSat 101: Basic Concepts and Processes for First-Time CubeSat Developers*. 2017. [Online] Available: https://www.nasa.gov/sites/default/files/atoms/files/nasa_csli_cubesat_101_508.pdf

[29] Cal Poly SLO, *CubeSat Design Specification (CDS) REV 13*, 2014. [Online]. Available: https://blogs.esa.int/philab/files/2019/11/RD-02_CubeSat_Design_Specification_Rev._13_The.pdf

[30] HQ AFSPC/SEK, *Air Force Space Command Manual 91-710, Volume 3*, May 2019. [Online]. Available: https://static.e-publishing.af.mil/production/1/afspc/publication/afspcman91-710v3/afspcman91-710v3.pdf

[31] C. Gebara and D. Spencer, "Verification and Validation Methods for the Prox-1 Mission," *Small Satell. Conf.*, Aug. 2016, [Online]. Available: https://digitalcommons.usu.edu/smallsat/2016/TS8StudentComp/3

[32] Jerry Buxton, *AMSAT Fox-1 DITL Test*. Jan 1, 2015 [Video Recording]. Available: https://www.youtube.com/watch?v=TjGAYvMyz4Q (accessed Dec. 31, 2020).

[33] Department of Defense, "MIL-STD-882E," May 2012. [Online]. Available: https://www.dau.edu/cop/armyesoh/DAU%20Sponsored%20Documents/MIL-STD-882E.pdf

[34] G. L. Prater, "NPSAT1 Missile System Pre-launch Safety Package (MSPSP)," Jun. 2004. Accessed: Jan. 1, 2021. [Online]. Available: https://apps.dtic.mil/sti/citations/ADA424941

[35] SpaceX, "Falcon User's Guide," Apr. 2020. [Online]. Available: https://www.spacex.com/media/falcon_users_guide_042020.pdf

[36] United Launch Alliance, *Delta IV Launch Services User's Guide*, Jun. 2013. [Online]. Available: https://www.ulalaunch.com/docs/default-source/rockets/delta-iv-user's-guide.pdf

[37] L. Valencia, "Autonomous Flight Termination System (AFTS)," 2019, [Online]. Available: https://www.gps.gov/cgsic/meetings/2019/valencia.pdf

[38] FCC. "Space Exploration Technologies Corp. (SpaceX) Experimental License FCC Filings." 2020. https://fcc.report/ELS/Space-Exploration-Technologies-Corp-SpaceX (accessed Jan. 2, 2021).

[39] Space Micro, "µSDR-C Software Defined Radio," 2019. [Online]. Available: spacemicro.com/products/communication-systems/µSDR-CTM%20SOFTWARE%20DEFINED%20RADIO.pdf

[40] Flexitech Aerospace. "Satellite Communication Systems Products." FlexitechAerospace.com. https://flexitechaerospace.com/products/ (accessed Jan. 2, 2021).

[41] CubeSatShop. "Helios deployable antenna." CubeSatShop.com. https://www.cubesatshop.com/product/helios-deployable-antenna/ (accessed Jan. 2, 2021).

[42] I. F. Akyildiz, J. M. Jornet, and S. Nie, "A new CubeSat design with reconfigurable multi-band radios for dynamic spectrum satellite communication networks," *Ad Hoc Nets.*, vol. 86, pp. 166–178, Apr. 2019, doi: 10.1016/j.adhoc.2018.12.004.

[43] E. Shehaj, V. Capuano, C. Botteron, P. Blunt, and P.-A. Farine, "GPS Based Navigation Performance Analysis within and beyond the Space Service Volume for Different Transmitters' Antenna Patterns," *Aerospace*, vol. 4, no. 3, Art. no. 3, Sep. 2017, doi: 10.3390/aerospace4030044.

[44] FCC, "GPS L1 Link Budget." [Online]. Available: https://apps.fcc.gov/els/GetAtt.html?id=110032&x=

[45] US Department of Transportation, "Global Positioning System (GPS) Adjacent Band Compatibility Assessment," 2017. [Online]. Available: https://www.transportation.gov/sites/dot.gov/files/docs/subdoc/186/dot-gps-adjacent-band-final-report.pdf

[46] ai-solutions, *FreeFlyer® Software*, v7.6 (Mission) [Commercial Software]. 2018. Available: https://ai-solutions.com/freeflyer-astrodynamic-software/

[47] W. Lan, R. Munakata, R. Nugent, and D. Pignatelli, "Poly Picosatellite Orbital Deployer Mk. III Rev. E User Guide," 2014. [Online]. Available: https://static1.squarespace.com/static/5418c831e4b0fa4ecac1bacd/t/5806854d6b8f5b8eb57b83bd/1476822350599/P-POD_MkIIIRevE_UserGuide_CP-PPODUG-1.0-1_Rev1.pdf

[48] MathWorks, *MATLAB Antenna Toolbox*, v2020b [Commercial Software]. 2020. Available: https://uk.mathworks.com/products/antenna.html

[49] Inside GNSS, "Measuring GNSS Signal Strength," *Inside GNSS - Global Navigation Satellite Systems Engineering, Policy, and Design*, Dec. 2, 2010. [Online]. Available: https://insidegnss.com/measuring-gnss-signal-strength/

[50] A. Brierley-Green, "Global Navigation Satellite System Fundamentals and Recent Advances in Receiver Design," presented at IEEE Long Island Section, Sep. 2017, [Online]. Available: https://www.ieee.li/pdf/viewgraphs/gnss_fundamentals.pdf

[51] G. X. Gao, M. Sgammini, M. Lu, and N. Kubo, "Protecting GNSS Receivers From Jamming and Interference," *Proc. IEEE*, vol. 104, no. 6, pp. 1327–1338, Jun. 2016, doi: 10.1109/JPROC.2016.2525938.

Possibilities and Limitations of Cyber Threat Intelligence in Energy Systems

Csaba Krasznay
Head of Institute
Institute of Cybersecurity
National University of Public Service
Budapest, Hungary
krasznay.csaba@uni-nke.hu

Gergő Gyebnár
Researcher
Black Cell Kft.
Budapest, Hungary
gergo.gyebnar@blackcell.hu

Abstract: The national energy system is the most critical of the critical infrastructures, and one which has become surprisingly vulnerable to cyberattacks in the last couple of years. Both unexpected technical design flaws and targeted attacks carried out by state-sponsored actors have raised challenges for the operators of essential services. Although this infrastructure is the subject of many regulations, and national security agencies pay special attention to such critical information infrastructures, gathering cyber threat intelligence is not straightforward for several reasons. First, special protocols in industrial control systems and operational technology (ICS/OT) systems are difficult to monitor. Second, information sharing does not really work, neither between states nor domestically. Third, due to the lack of thorough technical recommendations, there is no common understanding between responsible authorities and critical information infrastructure operators. In Hungary, key stakeholders of the national electricity system have realized that although some local and European legislation deals with the question of the cybersecurity of critical information infrastructure, many open questions remain in practice, both from policy and technology perspectives. In 2018, Hungarian manufacturers, energy service providers and responsible authorities started a discussion on what should be improved in legislation and technology, as well as in information sharing and how. This paper aims to describe the framework of this collaboration for information sharing and the initial results. Specifically, we present the current technical capabilities for gathering cyber threat intelligence in ICS/OT systems and propose some legislative actions that could support further technical solutions that are feasible in these special systems. We also present Tactics, Techniques, and Procedures (TTPs) and the goals of threat actors in energy systems that can be seen from the current data sets of our honeypots.

Moreover, we will also make some recommendations as to how the national and EU-wide legislation should be built up and what kinds of actions should be required from the key players in compliance with the Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union (NIS Directive).

Keywords: *ICS/OT security, energy cybersecurity, critical information infrastructure, NIS Directive, honeypot, ISAC*

1. INTRODUCTION

Energy is the most critical of the critical infrastructures. Without reliable energy services, our economy and society cannot operate. Related infrastructure has been attacked intensively from cyberspace since information technology became an inherent element of energy production and transmission. Most of the special systems were designed with safety in mind but not from a cybersecurity point of view, and therefore, as these industrial control systems and operational technology (ICS/OT) systems became interconnected, their built-in vulnerabilities were exposed to highly capable attackers who have sufficient knowledge to exploit them and who were state-sponsored. Moreover, due to the changing nature of energy consumption and the need for environment-friendly energy production, the whole industry has entered a paradigm shift, which involves currently unpredictable threats in the next decade.

As a result of these developments, the protection of critical information infrastructures has become a key concern for legislators, diplomats, and military leaders. According to Healey and Jankins, a cyberattack against the electric grid falls into the "Destabilizing Presence" category, which might invoke a direct answer from a country. [1] The European Union expressed the need for a joint diplomatic response to malicious cyber activities under the Cyber Diplomacy Toolbox, as the Council "expressed concerns about the increased ability and willingness of State and non-State actors to pursue their objectives by undertaking malicious cyber activities," by defining that "Cyber-attacks constituting a threat to Member States include those affecting information systems relating to, inter alia: (…) services necessary for the maintenance of essential social and/or economic activities, in particular in the sectors of: energy (electricity, oil and gas)." According to the Cyber Diplomacy Toolbox, "The Council stressed that clearly signalling the likely consequences of a joint Union diplomatic response to such malicious cyber activities influences the behavior of potential aggressors in cyberspace, thereby reinforcing the security of the Union and its Member States." [2]

The Directive on the security of network and information systems (NIS Directive) identifies the key types of entities related to the energy sector, or more precisely, the electricity system as essential services, in its Annex II:

- Electricity undertakings as defined in point (35) of Article 2 of Directive 2009/72/EC of the European Parliament and of the Council (1), which carry out the function of "supply" as defined in point (19) of Article 2 of that Directive;
- Distribution system operators as defined in point (6) of Article 2 of Directive 2009/72/EC;
- Transmission system operators as defined in point (4) of Article 2 of Directive 2009/72/EC. [3]

In practice, these declarations and legal texts could not achieve their goals without the extensive cooperation of the responsible national players as identified in the NIS Directive: the responsible national authorities, local ICS/OT and cybersecurity developers and service providers. In Hungary, the Security for Control Systems (SeConSys) initiative was established in 2018 to support the cooperation of these actors and facilitate the implementation of the NIS Directive, while increasing the competitiveness of Hungarian developers on the European market by providing leading cybersecurity technologies for the energy sector. Among others, the National Cyber Security Centre, which is designated as the National Single Point of Contact (SPOC) and acts as the national Computer Security Incident Response Team (CSIRT), as well as the sectoral authority – responsible for the designation of critical infrastructures in the energy subsector – are also part of this cooperation as can be seen in Figure 1. There are two working groups in SeConSys: one is responsible for regulatory questions, the other deals with technical challenges and both aim to provide an acceptable and feasible cybersecurity framework for the national electricity systems in compliance with the NIS Directive. As a result of this cooperation and with the support of the National Cyber Security Centre of Hungary, by the end of 2020, a Cyber Security Handbook for Electrical Industrial Control Systems was released and made publicly available.

FIGURE 1: MEMBERS OF SECURITY FOR CONTROL SYSTEMS (SECONSYS)

As the Handbook states in its chapter about the practical cyber defense of electricity systems,

> The operations management of the electricity system is a continuous, real-time process. The peculiarity of electricity is that the state of the system reacts very quickly to the control. The balance between consumption and production must be ensured under the right voltage conditions and the smooth running of business processes; all through the cooperation of many actors (across countries). Their feasibility today – and increasingly in the future – has made the operation of the electricity system dependent on ICS/SCADA components. The functionality of ICS/SCADA itself also depends on the power system. Although this chapter primarily makes recommendations for the IT/ICT sector, in line with the SeConSys approach, proper knowledge and consideration of OT specificities will also be provided. IT/ICT and OT security are valid together – the two areas need to be addressed together. In some cases, modifying an OT process makes the system as a whole less vulnerable from an ICT perspective, and special attention must be paid to ICT security for critical OT processes. In addition, due to the multi-stakeholder and geographically extensive connections, the system can be considered distributed and there is no complete control over it from any of the actors. [4]

The first recommendation of the Handbook stresses the importance of information sharing and gathering threat intelligence, in accordance with the feedback from the SeConSys members. The purpose of cyber threat intelligence is to provide background

information to enable management personnel to make informed decisions. This puts cyber security incidents in an appropriate professional context and supports hypothesis generation as a source at the beginning of incident management. In addition, it provides an opportunity for developing appropriate reactive defensive capabilities in relation to a specific event or sequence of events. Industry-specific reporting is essential for strategic (security management, organizational management), tactical (security teams, network teams, incident management teams) and operational (threat hunters, incident management teams, security management) organizations. This approach is aligned with Hungary's National Energy Strategy 2030, with an Outlook until 2040. The Strategy's declaration on cybersecurity highlights four action points: the creation of a sectoral recommendation (which is embodied by the Handbook), sectoral cyber threat information sharing, setting up a rapid incident management team and capacity building with skilled experts. [5]

As a widely accepted solution for cyber threat information sharing, in accordance with the relevant Hungarian strategies and other related legislation, the Hungarian Energy and Public Utility Regulatory Authority decided to establish an Information Sharing and Analysis Centre (ISAC) for the sectoral stakeholders who are also members of SeConSys. This body, known as E-ISAC, began operating in 2018. Below, we present our technical experiences on the collection and sharing of sector-specific cyber threat information for key stakeholders.

2. EXPERIENCES WITH ICS/OT CYBER THREAT INTELLIGENCE

When we set our goals in 2018, we decided to build a proper industry-specific cyber threat intelligence (CTI) feed for ICS/OT networks with a special focus on electricity. The reason why we chose this area is that the concept of Industry 4.0 may bring automation and comfort via the internet, but it also entails a huge risk for these devices. A myriad of threat feeds is available, but if they are not used properly, they can generate large quantities of noise and a slew of false positives. Moreover, these feeds are either too generic or do not cover some geographic locations properly. To avoid inadequate feeds, we decided to build an energy sector-specific honeypot network with sufficient territorial coverage that emulates the relevant protocols used by the industry.

First of all, it was important to note that there are some existing software applications for emulating ICS/OT protocols. However, the information derived from these is limited and does not meet our predefined requirement for the threat feed. In our concept, the threat feed should consist of a narrow layer of indicators of compromise (IoCs) and

other relevant repository-based rules that can be used for security operations (SecOps) in the field of threat hunting in ICS/OT infrastructures. We examined and tested the Conpot, GasPot, T-Pot, Dionaea, OpenPLC and MiniCPS frameworks. While all of these had advantages and disadvantages, we concluded that the best option for us was to develop our own software. First, we finalized the minimum viable product (MVP) protocol stack that represented the widely used protocols in the energy sector. These were Modbus, S7comm, IEC104 and generic IT protocols like telnet, ssh, http, and ftp. Other protocols, such as S7comm+, IEC101, IEC103 and IEC 61850 are to be included in a later phase as they were not identified as currently vital by the stakeholders. The second step was to define the level of interactivity. To leverage the power of CTI to effectively detect and respond to ICS related cyberattacks, it was clear that we needed to define the proper Tactics, Techniques, and Procedures (TTPs). Therefore, we used a map of TTPs based on the MITRE ICS ATT&CK framework, which "is a knowledge base useful for describing the actions an adversary may take while operating within an ICS network." [6] We plan to implement automatic support for the mapping of network data to MITRE ICS ATT&CK in the near future.

Initially, over 100 honeypots were virtually deployed in multiple cloud vendors. This was unsuccessful because it was not possible to simulate the real-life operation of such systems and adversaries could easily recognize that these are our honeypots. Subsequently, the number of our honeypots was reduced to 36 and then gradually increased to over 100. While carrying out this work, we realized that the design and implementation of honeypots for ICS is quite difficult on the infrastructure of cloud solution providers.

The major disadvantage of low-interaction honeypots is that they can easily be identified as decoys and thus cannot be used to examine the behavior of adversaries. However, the development and maintenance of high-interaction honeypots is challenging. To address these limitations, we decided to design a virtual, medium-interaction and server-side ICS honeypot that can be managed by a Software-Defined Network (SDN) controller using proxies. Our assumption was that such honeypots accessible over the internet are able to mimic a vulnerable interface that could determine the attackers' strategy. A broad spectrum of interactions is likely, including Denial-of-Service (DoS, flood the network), Man-in-the-Middle (MiTM) attacks, and device impersonation, which involves sending valid and malformed packets and the option of sabotage to trigger actions through malicious commands.

The following aspects were considered during the development of the infrastructure:
- Designing distributed and functionally separate elements;
- Using encrypted data connections between areas (e.g., VPS) and internal zones;

- Separating and protecting zones;
- High-speed data connections with minimal overheads;
- Simplifying the deployment of sensor devices;
- Minimizing maintenance needs for sensor devices (e.g., upgrades, configuration, new components);
- Monitoring and control of the condition of the sensors;
- Disconnecting sensor functions from actual VPSs, importing functions into the internal zone;
- Separating the processing zone from the zone containing the sensor functions;
- Creating a packet capture option;
- Grouping and virtualizing sensor functions (docker).

Due to the sensitive nature of this operational environment, further technical details cannot be shared. However, it is worth noting that our findings are similar to what Dodson, Beresford and Vingaard published in their paper [7]. Our goal was to validate and extend their results, which is why this paper does not examine other relevant ICS honeypot-related research. We can confirm that ICS/OT honeypots should be dispersed geographically, should be hosted on realistic IP addresses and not on cloud providers, should be high-interaction, and should be systematic and continuous. In order to gain better results, we recommend cooperation and information sharing between such honeypot operators, at least inside the European Union, in accordance with the requirements of the planned NIS2 Directive.

3. RESULTS

Our honeypots have been up and running since 2018. In order to measure and evaluate the success of their operation, we will review the data from our system between 1 November 2019 and 4 December 2020. This data set represents not just the number of attacks but also the history of the honeypot development. In that sense "attack" represents all successful interactions with the honeypots. We filtered out all mass scans and typical opportunistic nmap scans. At this stage, we were not able to distinguish between human and automatic bot-like activities. The reason for fluctuation stems from the availability of cloud providers, and the difference between the number of IT and ICS attacks can be explained by our initial lack of experience regarding ICS/OT knowledge. Our results are described in Table I and are explained below.

TABLE I: NUMBER OF IT AND ICS ATTACKS IN A GIVEN TIMEFRAME, DETECTED BY THE HONEYPOT SYSTEM

Interval start	Interval end	Number of IT attacks	telnet	http	ftp	dos	Number of ICS attacks	Modbus	S7comm	IEC104
2019.11.01	2019.12.01	949 898	949 898	-	-	-	-	-	-	-
2019.12.01	2020.01.01	5 178 366	5 178 352	14	-	-	27 736	27 736	-	-
2020.01.01	2020.02.01	5 677 315	5 677 269	11	-	35	37 998	37 998	-	-
2020.02.01	2020.03.01	11 320 234	11 300 972	10	8	19 244	17 653	17 653	-	-
2020.03.01	2020.04.01	5 056 354	5 050 695	2	-	5 657	22 948	22 948	-	-
2020.04.01	2020.05.01	2 429 267	2 425 523	1	-	3 743	17 257	17 257	-	-
2020.05.01	2020.06.01	88 315	88 022	-	-	293	755	755	-	-
2020.06.01	2020.07.01	2 429 813	2 427 785	2	-	2 026	7 731	7 731	-	-
2020.07.01	2020.08.01	1 317 754	1 316 275	5	1	1 473	10 944	10 944	-	-
2020.08.01	2020.09.01	200 656	200 416	-	-	240	1 451	1 451	-	-
2020.09.01	2020.10.01	84 544	70 287	13 058	930	-	26 524	13 059	12 518	947
2020.10.01	2020.11.01	131 168	107 888	21 788	1 226	-	1 558	260	-	1 298
2020.11.01	2020.12.01	66 654	43 360	17 530	955	-	267	267	-	-
2020.12.01	2020.12.04	6 308	4 138	1 599	131	-	18	18	-	-
SUM		34 930 338	34 840 880	54 020	3 251	32 711	172 840	158 077	12 518	2 245

Interval start: The start date of the measured data.

Interval end: The end date of the measured data.

Number of IT attacks: The aggregated attacks against emulated generic IT protocols, proxies, and environments.

http: The emulated webpages impersonate the web admin and login pages of Siemens and Moxa devices. Typical attack types detected were flooding, brute forcing, as well as a very small number of crafted/malformed HTTP packets.

telnet: Most of the attacks came from this source in proportion. Using a simple telnet emulation, we collected over 3 million unique IP addresses that were not previously recognized as bots. Most adversaries tried to block serial COM, while the rest tried to determine what information is shared between connected devices, including the particular hardware or software model. In some cases, approximately 6% of the adversaries tried to exploit known vulnerabilities associated with the protocol. In most cases, however, we experienced brute-force attacks. It should be highlighted that in February 2020 we detected an enormous number of attacks, double in numbers compared to the previous and the following month. This trend was also reported by various industry sources. For example, Microsoft Digital Defense Report stated that "IoT threats are constantly expanding and evolving. The first half of 2020 saw an approximate 35% increase in total attack volume compared to the second half of 2019." [8]

ftp: We set up an ftp server, which was used for sandboxing, with a user/password that could be easily guessed; for example, by using rockyou.txt, which is widely used by the users of Kali Linux as a default password dictionary. We assumed that the typical attacker would use Kali in that scenario. Sandboxing was implemented by our own static malware lab. In this period, we created 67 new YARA rules based on the examined IoCs that we had found in the uploaded content and shared these with the community. YARA is a widely used tool by malware researchers to identify and classify malware samples.

Number of ICS attacks: The aggregated attacks against emulated ICS/OT protocols and environments.

Modbus: Adversaries tried to establish command and control capabilities over Modbus to read the contents of the packets. They were looking for the IP address of the building management system (BMS) interface and the IP address of the receiving Modbus device to see the Function Code of the request. With all this data, the Modbus device became easily identifiable, and its Modbus Register Map revealed its control and command options. As soon as they had identified the device and its control commands via Modbus, there was no limit to further actions apart from the sandbox boundaries because they could simply begin to issue commands as though they were the BMS.

S7comm: Attackers conducted information gathering using the S7ReadArea, which allowed them to accurately map variables on the PLC, and then attempt to modify the variables; for example, by setting the request time for the modification fairly low, mostly lower than 20 milliseconds, allowing themselves to continuously overwrite it with specific values. This may cause unexpected behavior on the PLC. We also experienced some MiTM attacks.

IEC104: This widely used protocol had just a few hits, mostly from DoS and MiTM attacks, but in a very few cases we experienced unauthorized access to the input modules, the processor and the output. The attacks on the DoS were IEC104 packet flooding attacks. This attack type is kind of a DoS which aims to flood the Master Terminal Unit (MTU) with specific IEC104 command packets in order to generate a possible malfunction by the MTU. It confuses the system operator or even disrupts its operation. In the MiTM IEC 60870-5-104 isolation attack, the attackers aimed to isolate and drop the IEC104 network traffic between PLC and MTU. They performed an ARP poisoning attack utilizing Ettercap software, where a specific filter is widely available which isolates and drops the IEC104 packets between the PLC and MTU.

In most cases, connections came from bots or Mass Scan-like tools (78%) from already known malicious IP addresses. ICS/OT specific search engines like Shodan and Censys were the source of 13% of the connections, while 9% of the attacks came from previously unknown IP addresses. Table II illustrates the number of initiated connections toward our honeypots between August and December 2020. Each row represents a different IP address with different decoys in different regions. The numbers are relatively consistent, meaning that if the IP address of a potentially vulnerable ICS/OT system is revealed, it will be attacked immediately and continuously. It is also notable that the number of ICS/OT targeting attacks is significantly lower than the number of IT attacks. We assume that ICS/OT knowledge is still owned by a minority of cyberattackers; therefore, companies operating special protocols should be prepared for highly skilled attackers as adversaries.

TABLE II: NUMBER OF DETECTED ATTACKS ON DIFFERENT IP ADDRESSES

Country	Number of connections
India	224 193
Singapore	179 132
India	177 674
Netherlands	175 710
Germany	171 926
Germany	171 659
Germany	171 000
Netherlands	170 941
Germany	170 330
Singapore	169 649
United Kingdom	169 621
Singapore	169 133
Germany	169 053
United Kingdom	168 131
Germany	167 219
Singapore	166 716
Singapore	166 275
Singapore	164 087
India	152 647

United Kingdom	151 726
Germany	150 122
Germany	129 184
Germany	123 769
United Kingdom	123 434
Germany	122 729
Netherlands	121 895
Netherlands	120 677
Netherlands	118 850
Germany	108 215
United Kingdom	99 042
United Kingdom	96 254
United Kingdom	95 098
Singapore	11 864
United Kingdom	9 130
Singapore	6 429

4. USING CYBER THREAT INTELLIGENCE IN PRACTICE

The latest Cyber threat intelligence overview prepared by the European Union Agency for Cyber Security (ENISA) summarizes the major requirements for CTI as follows:

- Cooperation and coordination of EU-wide CTI activities;
- Identification of CTI requirements;
- Facilitation of CTI's connection with geopolitical information and cyber-physical systems;
- Integrating CTI with security management processes;
- Development of a comprehensive CTI program by ENISA;
- Investment in some basic CTI concepts, in particular CTI maturity and threat hierarchies.

This overview also contains the results of a comprehensive CTI survey conducted by ENISA of interested stakeholders. The survey highlights current trends relating

to the way in which CTI is managed from practical and technical perspectives – the following includes excerpts from the report:

- Semi-automation of CTI production is an important tool, but manual activities continue to comprise the core of CTI production;
- Information aggregation, analysis and dissemination activities are managed using widely available tools such as spreadsheets, mail and open-source management platforms, which is indicative of the efficiency of low-cost solutions;
- The importance of defining CTI requirements is understood by the CTI user-community – this is an indication that CTI is becoming part of decision-making at business and management levels;
- A combination of consumption and production of CTI is the prevailing method for building up an internal CTI knowledge base;
- Open-source information gathering is the most widely used ingestion method, followed by threat feeds from CTI vendors;
- Threat detection is assessed as the main use case for CTI; although indicators of compromise (IoCs) are still the most important elements of CTI in threat detection and threat response, threat behavior and adversary tactics (TTPs), seem to be responsible for the upwards trends in the use of CTI in organizations;
- Measuring the effectiveness of CTI is still a difficult task. An interesting finding regarding the level of satisfaction is the low rating given to the value of machine learning functions. [9]

In general, we can confirm these findings based on our experience. We wish to emphasize the importance of understanding TTPs from the list above. This allows us to understand the techniques and procedures and to link an attack, for example, to the MITRE ICS ATT&CK framework, which represents a useful knowledge base for describing the actions an adversary may take while operating within an ICS network. This kind of knowledge base can also be used to better characterize and describe post-compromise adversary behavior. In contrast to the results of ENISA's survey, we obtained promising preliminary results with machine learning-based predictions, and these may be the subject of a future paper. We assume that better and more extensive knowledge of machine learning, or artificial intelligence more generally could increase the efficiency of the everyday usage of such technologies in cyber threat intelligence.

To illustrate the importance of understanding TTPs, we will outline a cyber incident that has not yet been published. In this case, a financial investigation found that somebody had earned millions of dollars in a short transaction on an energy company. The investigation was successful and found that the attackers had downgraded and

synchronized all the protection relays, stopping the relays from working at a given time. This resulted in a serious loss in both production and share value.

The adversary's tactic was to inhibit the response function. It achieved this by modifying the control logic, using procedures very similar to Triton malware. This could be determined because the Human-Machine Interface (HMI) registry logs had been parsed to a Security Incident and Event Management (SIEM) system and there was a correlation rule with the proper ICS threat feed that contained Triton's registry key modifications. Solely gathering IoCs would not have been enough. We needed to put these IoCs in context and had to have workflows, implemented and tuned use cases, threat hunting, triage, and other proactive workflows.

Besides the information security aspects of the above-mentioned cyberattack, such information would also be very valuable for the local authorities. As has been known since 2017, Triton is actively targeting ICS systems. One of the earliest warnings came from FireEye. Their threat research report clearly describes relevant IoCs, but their speculations on the intent of the attacks remain within the targeted organization. The research paper claims that, "We assess with moderate confidence that the attacker's long-term objective was to develop the capability to cause a physical consequence. We base this on the fact that the attacker initially obtained a reliable foothold on the DCS and could have developed the capability to manipulate the process or shutdown the plant, but instead proceeded to compromise the SIS system. Compromising both the DCS and SIS system would enable the attacker to develop and carry out an attack that causes the maximum amount of damage allowed by the physical and mechanical safeguards in place." [10] There is no mention of any financial intent. Moreover, in October 2020, the U.S. Department of Treasury announced sanctions against the State Research Center of the Russian Federation FGUP Central Scientific Research Institute of Chemistry and Mechanics (TsNIIKhM), a Russian government-controlled research institution, which was attributed as a responsible party for building the customized tools that enabled the Triton attack. The reasoning is that "researchers who investigated the cyber-attack and the malware reported that Triton was designed to give the attackers complete control of infected systems and had the capability to cause significant physical damage and loss of life." In this case, financial motivation was not mentioned either. [11]

Our major argument for cyber threat intelligence information sharing is that if local and European authorities had the relevant information on the "dual-use" of Triton (meaning to earn money and not "only" to prepare for physical destruction) and they shared this information with private companies who might be potential victims, a higher level of cyber preparedness would be achieved. We assume that potential financial loss is a higher motivation than a potential outage. Moreover, we assume

that financial gain derived by the cyberattackers would finance other illicit operations in the future. If Western countries could cut off such illegal income streams from these allegedly state-sponsored groups, their operational capabilities would be lowered.

We believe that the capability of processing such CTI requires a higher level of cybersecurity maturity on the part of the organizations targeted. Therefore, we recommend that the organizations conduct self-assessments before the implementation of CTI. Predefined maturity frameworks of this type have been published by many organizations. We suggest using the Cybersecurity Maturity Model Certification (CMMC) developed by Carnegie Mellon University and Johns Hopkins University. According to CMMC, organizations at Level 3 are mature enough to "receive and respond to cyber threat intelligence from information sharing forums and sources and communicate to stakeholders." [12]

To describe an incident, we recommend using Structured Threat Information Expression (STIX), version 2.1, "that is a language and serialization format used to exchange cyber threat intelligence (CTI)." [13] We created our CTI feed using the standardized methods of STIX 2.1. We share this information via the Trusted Automated Exchange of Intelligence Information (TAXII), which is an application protocol for exchanging CTI over HTTPS. "TAXII defines a RESTful API (a set of services and message exchanges) and a set of requirements for TAXII Clients and Servers." [14] We not only collect IoCs but also correlate them into context using external feeds for better triage for SecOps.

5. RECOMMENDED STEPS TO DEVELOP, EXPAND AND ENHANCE ICS/OT THREAT INTELLIGENCE

Of key importance for gaining relevant feeds and context is the power of the sector-specific crowdsource. The best option is to give ISACs the ability to act as a threat intelligence platform (TIP). The importance of ISACs will increase with the rise of information technology, Industry 4.0 and 5.0. Their goal is to respond to the cybersecurity challenges generated within the industry by bringing the stakeholders together on a centralized platform. An ISAC must meet both human-to-human and machine-to-machine needs. Accordingly, traditionally accepted "human-readable intelligence" functions are no longer sufficient. Next-generation ISACs must harmonize knowledge that can be processed, shared, and distributed by both human and machine means, by hosting repository-based servers such as the Malware Information Sharing Platform (MISP) or TAXII. This ability is not tomorrow's technology, but yesterday's competition, with the advent of machine-to-machine AI-based attacks and defense, where manual human interaction is not enough. Therefore, these ISACs have to have

two main scopes: human-readable intelligence and repository-based intelligence. We propose the structure below for information sharing on an ISAC platform. This structure was implemented on the Hungarian E-ISAC, and as such has been tested in a real-life environment.

Human Readable Intelligence

Our ISAC framework includes some basic ISAC functions that enable the whole sector or just one entity to use it as a "virtual war room" defense communication platform in case of a coordinated cyberattack. This functionality can support situational awareness. Human readable intelligence can be likened to a social media platform, such as Twitter, that informs the user about relevant cases in a predefined scope, which a stakeholder can "follow." Specific newsletters and vulnerability disclosures are also part of this threat intelligence feed. We provide the following sections for the stakeholders.

- *Report an incident*
  - *Anonymity:* The tab allows both anonymous and named incident reporting for authorized users. Anonymity is important because market competition within the sector can override information sharing, making the whole crowdsourcing project ineffective.
  - *Ticketing:* Incidents can be integrated with most ticketing tools (JIRA, SNOW, etc.), and the platform can also send email and SMS notifications directly.
- *Forum*
  - The forum serves to share upcoming tasks, sector-specific problems and solutions.
- *Documents*
  - Uploaded documents with descriptions of them are collected under Documents. Various categories, file visibility and permissions can be set individually.
- *News*
  - A classic news thread with many administrative and aggregation options.
- *Events*
  - Reminders and announced events can be published (Exercises, Expos, conferences, TTX, Range / Drill, etc.) The iCal function can be used to save the selected event to the user's calendar. Only the site administrator has permission to announce an event.

- *Site feed news*
  - Information about the collected resources (TTPs, Tools, Campaigns, Alerts, IoCs, etc.) as well as their distribution by type is found on this page.
- *Incident response*
  - If a dedicated CSIRT / CERT is available to the sector, then the entity's direct, dedicated contact details are displayed here.

Repository-based Threat Intelligence

One of the goals of these ISACs is to broadcast and spread the threat feed, which we achieve using integrated solutions such as MISP, STIX or TAXII. With the help of the technology, the organization and the entire sector can automate the detection of IoCs identified while hunting for threats. Furthermore, stakeholders can jointly perform malware analysis. Through this crowdsource power, the strength of the community can leverage the repository-based threat intelligence SecOps activity via tools like CybOx – integrated into STIX 2.0 – where the community can work together on malware analysis or even on a cyber kill chain. Such repository-based threat intelligence could also be used for other SecOps activities to feed SIEM, Security Orchestration, Automation and Response (SOAR), Intrusion Detection and Preventions Systems (IDPS) and threat hunting platforms, in the same way that antivirus or IDPS vendors do. The distribution of threat feeds is the privilege of the umbrella organization (for example the sectoral ISAC).

The platform should employ a sector-specific, deception-based intrusion detection infrastructure that enriches incoming data with relevant context (domain information, IP information, malware hash, botnet vulnerability database, etc.). We recommend the members of each ISAC produce sector-specific feeds. That requires a customized decoy and honeynet infrastructure, including DNS Honeypot, Honeytokens, ICS honeypots and honey personas.

6. CONCLUSION

In Hungary, the Hungarian Energy and Public Utility Regulatory Authority decided to set up an energy sector-specific ISAC, called E-ISAC, in accordance with Hungarian and local strategies and legislation. Its aim was to implement both human-readable intelligence and repository-based intelligence. However, during the implementation phase, we realized that there are no exact technical requirements or recommendations on how to implement the information sharing platform of E-ISAC. Neither the Authority nor the participants provided clear technical specifications. According to ENISA's report on NIS investments, this is a common problem in Europe:

Irrespective of organizations' current implementation state, the challenges that were most cited were the prioritization of other regulations, the existence of stronger local regulations and the lack of clarity of the NIS Directive expectations after transposition into national law. However, for organizations that do not have a dedicated NIS Directive implementation project, internal challenges such as the lack of resources (34.6% of such respondents), lack of skills (30.8%) and lack of collaboration (30.8%) appear to be most important. [15]

As the new EU Cybersecurity Strategy for the Digital Decade, released in December 2020, states,

The Commission proposes to build a network of Security Operations Centres across the EU, and to support the improvement of existing centres and the establishment of new ones. (…) The centres would then be able to more efficiently share and correlate the signals detected and create high-quality threat intelligence to be shared with ISACs and national authorities, and thus enabling a fuller situational awareness. [16]

Based on our experience, we recommend the establishment of a European-wide, clear technical standard for cyber threat information sharing. We believe that the Strategy's goal ("to connect, in phases, as many centres as possible across the EU to create collective knowledge and share best practices") cannot be achieved without standardization. Therefore, we propose that ENISA and the European Telecommunications Standards Institute (ETSI) create a new European standard for cyber threat information sharing, based on the widely used STIX and TAXII protocols. We also recommend that the European Commission refer to this standard in the revised NIS Directive or "NIS 2" as a mandatory requirement for member states and organizations under the NIS Directive. Moreover, we recommend the creation of a threat intelligence feed with limited access for NIS obliged organizations at least on the national level. Such a feed could be financed by governments or by organizations through obligatory ISAC membership. CTI is the first step toward early warning and successful defense in cyberspace over the next decade. This is a basic requirement for SecOps and provides a unique opportunity for threat hunting for both private companies and national security authorities.

REFERENCES

[1] J. Healey and N. Jenkins, "Rough-and-Ready: A Policy Framework to Determine if Cyber Deterrence is Working or Failing," in *11th International Conference on Cyber Conflict: Silent Battle*, T. Minárik, S. Alatalu, S. Biondi, M. Signoretti, I. Tolga and G. Visky, Eds., Tallinn, NATO CCD COE Publications, 2019, pp. 123–142.

[2] Council of the European Union, "COUNCIL DECISION concerning restrictive measures against cyber-attacks threatening the Union or its Member States," 14 May 2019. [Online]. Available: http://data.consilium.europa.eu/doc/document/ST-7299-2019-INIT/en/pdf. [Accessed 29 December 2020].

[3] *Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union*, 2016.

[4] P. Görgey and C. Krasznay, Villamosenergetikai ipari felügyeleti rendszerek kiberbiztonsági kézikönyve, Budapest: Nemzetbiztonsági Szakszolgálat Nemzeti Kibervédelmi Intézet, 2020.

[5] *Nemzeti Energiastratégia 2030, kitekintéssel 2040-ig*, 2020.

[6] The MITRE Corporation, "ATT&CK® for Industrial Control Systems," The MITRE Corporation, 3 June 2020. [Online]. Available: https://collaborate.mitre.org/attackics/index.php/Main_Page. [Accessed 29 December 2020].

[7] M. Dodson, A. R. Beresford and M. Vingaard, "Using Global Honeypot Networks to Detect Targeted ICS Attacks," in *12th International Conference on Cyber Conflict - 20/20 Vision: The Next Decade*, T. Jančárková, L. Lindström, M. Signoretti, I. Tolga and G. Visky, Eds., Tallinn, NATO CCDCOE Publications, 2020, pp. 275–291.

[8] T. Burt, "Microsoft report shows increasing sophistication of cyber threats," 29 September 2020. [Online]. Available: https://blogs.microsoft.com/on-the-issues/2020/09/29/microsoft-digital-defense-report-cyber-threats/. [Accessed 2 January 2021].

[9] European Union Agency for Cybersecurity (ENISA), "Cyberthreat intelligence overview," European Union Agency for Cybersecurity (ENISA), Attiki, Greece, 2020.

[10] B. Johnson, D. Caban, M. Krotofil, D. Scali, N. Brubaker and C. Glyer, "Attackers Deploy New ICS Attack Framework "TRITON" and Cause Operational Disruption to Critical Infrastructure," 14 December 2017. [Online]. Available: https://www.fireeye.com/blog/threat-research/2017/12/attackers-deploy-new-ics-attack-framework-triton.html. [Accessed 27 December 2020].

[11] U.S. Department of the Treasury, "Treasury Sanctions Russian Government Research Institution Connected to the Triton Malware," 23 October 2020. [Online]. Available: https://home.treasury.gov/news/press-releases/sm1162. [Accessed 3 January 2021].

[12] Carnegie Mellon University and The Johns Hopkins University Applied Physics Laboratory LLC, "Cybersecurity Maturity Model Certification (CMMC)," 18 March 2020. [Online]. Available: https://www.acq.osd.mil/cmmc/docs/CMMC_ModelMain_V1.02_20200318.pdf. [Accessed 28 December 2020].

[13] Cyber Threat Intelligence Technical Committee, "Introduction to STIX," 29 November 2020. [Online]. Available: https://oasis-open.github.io/cti-documentation/stix/intro.html. [Accessed 2 January 2020].

[14] Cyber Threat Intelligence Technical Committee, "Introduction to TAXII," 29 November 2020. [Online]. Available: https://oasis-open.github.io/cti-documentation/taxii/intro.html. [Accessed 3 January 2021].

[15] A. Drougkas, G. Bafoutsou and V. Paggio, "NIS Investments," European Union Agency for Cybersecurity, Heraklion, Greece, 2020.

[16] European Commission, "New EU Cybersecurity Strategy and new rules to make physical and digital critical entities more resilient," 16 December 2020. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2391. [Accessed 27 December 2020].

Building a National Cyber Strategy: The Process and Implications of the Cyberspace Solarium Commission Report

Brandon Valeriano
Donald Bren Chair of Military
Innovation
Marine Corps University
Quantico, VA, USA

Benjamin Jensen
Professor
School of Advanced Warfighting
Marine Corps University
Quantico, VA, USA

Abstract: Crafting a national cyber strategy is an enormous undertaking. In this article we review the process by which the Cyberspace Solarium Commission generated the Solarium Commission Report, developed the strategy of layered cyber deterrence, and strategized for legislative success in implementing its recommendations. This is an article about the development of a whole-of-nation strategy. Once the production of the strategy of layered cyber deterrence is explained, the article goes on to elaborate on implementation strategies, the challenge of escalation management, and future efforts to ensure that the work of the Solarium Commission becomes entrenched in U.S. national cyber strategy and behavior. We review the work left undone by the Solarium Commission, highlighting the enormous effort that went into the process of building out a strategy to defend a nation.[1]

Keywords: *cyber strategy, deterrence, coercion, escalation*

---

[1] It takes a village; we thank the entire Solarium Commission team, as their efforts generated the final Commission Report and the legislative successes that followed. In some ways, this article seeks to chronicle the process of building a strategy that was developed through the efforts of hundreds of people. This work reflects the process that we went through to construct the Solarium Commission report, which is particular to our experience; others may have had different recollections of the events under consideration. Brandon Valeriano is also a Senior Fellow at the Cato Institute and a Senior Advisor to the Cyberspace Solarium Commission. Benjamin Jensen is also a Scholar in Residence at American University and the Research Director for the Cyberspace Solarium Commission.

1. INTRODUCTION

Established in the fiscal year 2019, the John S. McCain National Defense Authorization Act (NDAA) created the Cyberspace Solarium Commission to evaluate competing approaches of cyber strategy and seek a consensus comprehensive strategy to defend the United States in cyberspace against significant attacks. This article will review the process of developing the report of the Cyberspace Solarium Commission (hereafter, Solarium Commission) (Montgomery, Jensen et al. 2020) and the strategy of layered cyber deterrence (Jensen 2020).

The challenge of "develop[ing] a consensus on a strategic approach" is immense (Congress 2017–2018, 132 STAT. 2141). The Fiscal Year 2019 NDAA tasked the Solarium Commission with considering the options of "deterrence, norms-based regimes, and active disruption of adversary attacks through persistent engagement" (Congress 2017–2018, 132 STAT. 2143). These options became overlapping layers, mimicking the original Eisenhower Solarium Commission's strategy of engagement with the Soviet Union combined with aspects of containment and deterrence (Gallagher 2015). Rather than viewing strategic approaches as mutually exclusive, the team viewed them as complementary, creating an overall denial-based effect on adversary decision-making.

The central idea behind layered cyber deterrence is to alter the cost-benefit calculations of the adversary to threaten U.S. interests in cyberspace yet also take into account the global reliance private sector networks have on the new digital commons. No action will stop all cyber activity by state and non-state actors engaged in political warfare, espionage, military operations, or criminal activity. Rather, the goal is to alter the cost-benefit calculation to reduce the severity and frequency of cyber activity.

The first layer became "shape behavior," encompassing the development of normative regimes to govern cyberspace in collaboration with international partnerships. Shaping behavior also seeks to leverage non-military instruments such as regulations and legal regimes to produce a cyber environment that favors stability. Entanglement (Nye 2017), another term for shaping the international environment, includes not only norm generation but also the inclusion of various structures that could facilitate progress in cyber security to shape the environment in ways that are conducive to global security. The second layer became "deny benefits," which encompasses some traditional aspects of deterrence but focuses on resiliency and defense in depth (Valeriano and Jensen 2019). This effort includes securing elections, protecting critical infrastructure, and ensuring the continuity of the economy and government. By hardening the defense targets, the U.S. can enable deterrence and forestall digital violence.

The third layer became "impose costs," which sought to generate cyber capabilities and capacity.[2] The goal was to flesh out the concept of persistent engagement (Fischerkeller and Harknett 2017; Healey 2019). Persistent engagement suggests that imposing costs was an outgrowth of the strategy, not the means (Fischerkeller and Harknett 2020). To orchestrate a whole-of-nation approach to defending the nation through forward action and cost imposition, the Solarium Commission recommended enabling the United States to leverage cyber power to achieve effects, but with an eye towards preserving privacy, the resilience of global networks, and the proper delegation of authorities, consistent with international law and existing legal regimes.

In this paper we review how the strategy of layered cyber deterrence was constructed and how the background research and wargames helped the Solarium Commission staff generate the final report (Montgomery, Jensen et al. 2020), released in March 2020. We will then evaluate the successes and the challenges of the Solarium Commission, highlighting potential criticisms and outlining a path forward as the Biden administration takes the reins in national policy.

Developing and implementing a strategy for defending a nation-state in cyberspace is a difficult proposition given all the agencies, interests, and fixed positions of those operating in the defense and cyber policy ecosystem. By valuing originality, empirical research and seeking to achieve a bipartisan goal of developing a comprehensive national strategy, the Solarium Commission Report is an example of a progressive method of generating a national strategy to defend the nation against adversaries. This article will explain the process by which the Solarium Commission strategy was built while also considering the challenge of escalation risk management.

2. THE CHALLENGE OF CREATING A NATIONAL CYBER STRATEGY

A. Building a National Strategy

There are few manuals on how to draft a national strategy. Academics tend to be better at judging other people's strategies than they are at developing organized, deliberative processes to generate policy recommendations and clear tasks for government agencies. Yet policymakers tend to see the domain of crafting strategy as – to paraphrase Hobbes – a nasty, brutish, and short battle of ideas rooted as much in gaining positional or transactional bureaucratic leverage as it is in analytical clarity and logical consistency (Jensen 2018).

---

[2] The Solarium Commission did not develop methods to impose costs on the adversary; instead, the task was to enable the U.S. government to able to impose costs by setting it up for action. This came in the form of enabling workforce development and strategic assessments within the DoD to providing recommendations for the evolution of the State Department.

Absent a guiding process to evaluate ideas and test assumptions, strategy formation devolves into a competition between competing bureaucratic interests. Policy entrepreneurs compromise in pursuit of an agenda (Kingdon and Stano 1984; Durant and Diehl 1989; Mintrom 1997). The result is a "garbage can" full of ideas – some good, others bad, many irrelevant to the problem at hand (Cohen, March et al. 1972). The process by which one develops a strategy is as important, if not more important, than the resulting blueprint for aligning limited resources in pursuit of national objectives, given fixed preferences and risk considerations (Klimburg 2012). A clear, deliberative process can guard against some of the agenda-setting dynamics as well as check other common sources of bias. The goal is to make the process transparent and open to periodic checks with a larger set of stakeholders. Careful attention to process and risk mitigation provides decision-makers with a venue for understanding their own preferences and inherent tradeoffs in any policy selected.

Building the Team of Strategists

Concern for creating a marketplace of ideas guided the early stages of building a team and process for the Solarium Commission. Starting in early spring 2019, a small team began to meet with Executive Director Mark Montgomery (retired rear admiral and former policy director for the Senate Armed Services Committee), and his chief of staff, Deborah Gray (retired colonel, U.S. Army), to develop a plan of action.

The NDAA had already specified the research lines of effort, and the Solarium Commission started deliberating, staffing the task force leads, and hiring support staff. Dr. Erica Borghard, an academic, led Task Force One. John Costello, an appointee detailed from the Department of Homeland Security (DHS), led Task Force Two. Val Colfield, a senior official from the FBI, led Task Force Three. Cory Simpson, a lawyer with recent experience at U.S. Cyber Command (USCYBERCOM), organized and led a general support element dubbed the Fourth Directorate. Dr. Benjamin Jensen served as senior research director and lead author, organizing the process to develop the strategy, crafting deliberative mechanisms including the Red Team and Solarium event, and creating the core strategy: layered cyber deterrence.[3]

Next, the Commission built out its staff at the direction of the Task Force leads and Commission members, interviewing and hiring Commission team staffers from Capitol Hill offices, think tanks, and academia. After key hires and detailed personnel were in place, the executive director, task force leads, and senior research director, to use military jargon, "planned the plan," mapping out a timeline, key deliverables, and the overarching process to evaluate each task force effort and to build the final strategy with the commissioners.

---

[3]   The full list of staff and contributors is accessible at https://www.solarium.gov/about/staff and in the Solarium Commission Report.

In spring 2019, the appointed members of the Commission began to meet for progress reviews. The executive director used these meetings to update the Commission on progress and timelines and to solicit any additional input. The general format was that the staff products were briefed to Commission members who then would follow up and consult individual teams on their activities, shaping the report and recommendations in collaboration between Commission staff and Commission members.

B. The Process of Building a Strategy

To initiate the strategic formation process, the senior research director built on the original Eisenhower Solarium effort. The purpose was not to carbon-copy the effort but to use it as a lens through which to develop a deliberative strategy formation process. The idea was to progress from task force research to Commission approval and ultimately legislative or executive branch action based on the proposed policies.

This effort started by briefing each task force on the original Solarium effort and illustrating how competitive teams in that process organized their reporting. The senior research director distributed declassified copies of the original Solarium reports and used them to work with task force leads on the structure of their submission. Figure 1 provides a sample product used during this phase, showing the task forces' different report structures and internal logics used in the 1953 effort.

FIGURE 1: TASK FORCE PLAN FOR EISENHOWER SOLARIUM

Task Force A July 1953
I. The Task
II. The Situation Before the United States
III. Courses of Action for U.S. Policy A. Maintenance of U.S. Strength B. Maintenance of the Economy C. Maintenance of Free Political Institutions D. Strengthening the Free World (general, Europe, Asia, Middle East, North Africa) E. Prevention of Soviet Expansion F. Reduction of Soviet Power G. Establishment of International Order
IV. Costs
V. Review of Conclusions in Light of Three Alternative Lines of Soviet Action
VI. Comments Regarding Questions in Section III.2 of Project Paper
VII. Summary and Concluding Statements

Task Force B July 1953
I. The Situation which the United States and the Free World Must Meet
II. Recommended National Policy A. Statement of the Policy B. Clarification of Policy C. Rejected Alternative Lines D. Rejection of the "Two Worlds" Concept
III. Summary of Advantages A. Enumeration of Principal Advantages B. Effectiveness of Alternative "B" in Meeting Possible Soviet Lines of Action
IV. Analysis of Implications A. Military Implications of Alternative "B" B. Political and Psychological Implications of Alternative "B" C. Economic Implications D. Probable Soviet Reactions to Alternative "B"
V. Weakness of Alternative "B" A. Introduction B. Soviet Capabilities and Intentions C. Effects of Other Free World Countries D. Support of the Policy by the American People
VI. Implementation A. General Considerations B. Specific Proposals
Enclosures 1. The Role of General War Under Alternative A, B, and C 2. Possible Implications of Measures
Annexes A. Examination of Alternative "B" in the Memorandum on Basic Issues B. U.S. Commitments in Regard to the Defense of Countries Subject to Armed Attack by the Soviet Bloc

Departing from the original Solarium, the Cyberspace Solarium effort opted to have each task force submit not just a strategic approach but a formal workplan organized around key questions. The reason for organizing around questions, as opposed to exclusively around policy approaches, was to ensure a more open research phase. While each task force used a common approach in the form of a workplan, the Fourth Directorate served more as general support. This group developed the threat assessment narrative and explored topics, like artificial intelligence and elections, that emerged during the research phase.

With the workplans in place, the teams initiated a compressed six-month process of conducting research and using the insights to refine their initial strategic approach and policy recommendations. During this time, the Commission held progress review meetings, in which the executive director would have various task force leads and staff brief key findings and initial perspectives based on their workplan. These meetings helped the Commission identify more contentious areas and collect additional concerns that would need to be addressed during the Solarium event. In addition, the executive director, Mark Montgomery, held a series of meetings with different Commission staff weekly to identify additional issues and concerns. It was not uncommon for Commission staff, especially the task force leads and senior research director, to meet privately with elected officials and senior appointees across government. To summarize this approach, the staff used the placemat in Figure 2 to aid in outlining the task force organization, logic, and timeline.

FIGURE 2: THE STRATEGY FORMATION PROCESS PLACEMAT
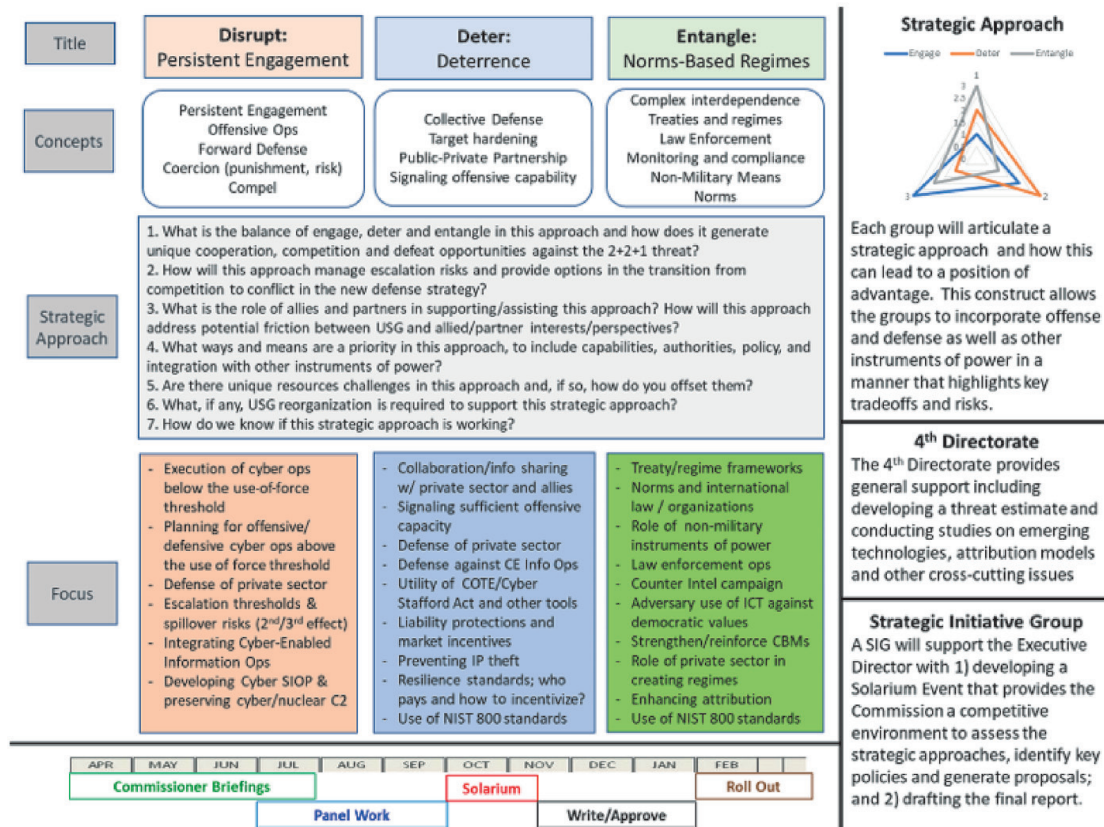
The research phase was extensive, involving over 300 interviews and reviewing over 30 submissions from academics and thought leaders in cyber security. The staff also traveled, attending meetings with officials involved in cyber policy and the private sector at events like the DEF CON hacking conference in Las Vegas and to Europe, with a particular focus on the United Kingdom, Estonia, and Israel. There were targeted trips with multiple events and meetings to San Francisco (Silicon Valley), New York City (financial sector), and Boston to consult with cyber security experts.

Towards the end of summer, the teams began to transition to panel work, essentially triaging the various answers they found through research to the core and derivative questions referenced in the workplan. The result was a task force strategy and linked policy recommendations. Each task force approached this phase slightly differently. Some took a more top-down approach, crafting ideas and then socializing them. Others divided their task force into teams focused on areas or worked each issue collaboratively. The executive director kept an open-door policy to hear any emerging concerns and used a weekly meeting to check progress. During these progress reviews, alongside the larger meetings with the Commission, the senior research

director worked to finalize the deliberative mechanisms the commissioners would use to evaluate each task force: 1) a Red Team and 2) the Solarium event.

FIGURE 3: THE SOLARIUM COMMISSION ROAD MAP
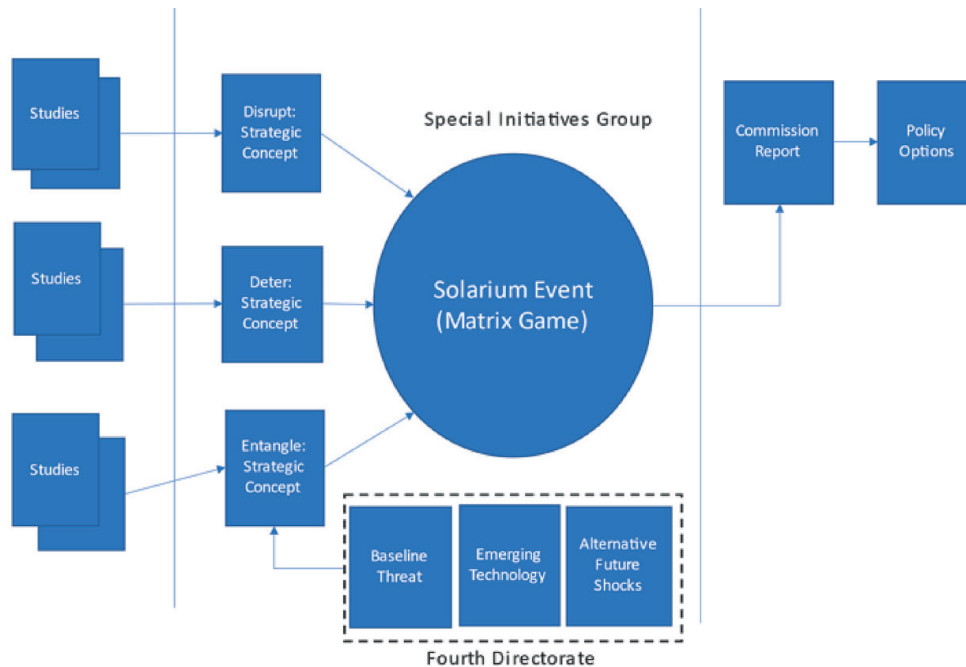
C. The Emerging Strategy and Solarium Event

From October 21 to 23, 2019, the task forces submitted their initial strategic approaches, based on their research and answers to the questions in the workplan, to a Red Team. Red teams are a common military, intelligence, and business community mechanism to identify critical assumptions and evaluate alternative perspectives by acting as the "enemy" (Zenko 2015). Applied to the Solarium Commission, the Red Team engaged predominantly in challenge activities, forcing each team to clarify their logic (e.g., theory of victory, principles) and the way policy recommendations related to core problems the task force identified. Members of the Red Team included retired flag officers, former senior National Security Council officials, and leading cyber experts from industry.[4] After the Red Team review, task forces used October 24 to prepare for the Solarium Event.

The Solarium event combined elements of red teaming, matrix wargames, and stress tests to create a deliberative environment for commissioners to evaluate each task force. The senior research director developed two scenarios linked to the baseline threat and issues previously identified by the commissioners. These scenarios, *Slow Burn* and *Break Glass*, used hypothetical countries and incorporated a wide range of both previously observed cyber incidents and more catastrophic possibilities. These

---

4   The complete Red Team list is available at: https://www.solarium.gov/.

scenarios, as stress tests, allowed the task forces to outline how their approach would do three things: 1) prevent the hypothetical cyber crisis; 2) provide options to respond to the cyber crisis; and 3) support the government and private sector in mitigating the consequences of the cyber crisis. As seen in Figure 4, the day was organized into four sessions. First, the baseline threat estimate was briefed to the commissioners. Second, the task forces responded to the first of two scenarios. The senior research director served as the moderator, ensuring each task force had an opportunity to outline its answers to the three questions. After the initial response, the Red Team asked questions and opened the floor to the commissioners for any follow-up questions.

The scenarios functioned as both stress tests and a modified matrix game. While there was no opponent per se, the task force leads had to account for how well their underlying strategy and linked recommendations would address the cyber crisis. While commissioners listened to the responses, they filled out their evaluation of the underlying policy recommendations using a rank-ordered system. Each commissioner privately rated each recommendation from 1 to 5, with 5 being the highest rating. They were also allowed to assign a relative weight, indicating how important the recommendation was to them. This system allowed the staff to identify areas of convergence and divergence between the commissioners. This approach proved critical in that it allowed the staff to quantitatively show the commissioners what they agreed on, thus maximizing time for debate in future meetings.

After the Solarium Event, the Commission deliberated from November to January. These sessions benefited from the ranked weighting system and the wide range of perspectives offered by the Red Team. Concurrently, the staff worked to narrow the range of recommendations (initially there were over 100 recommendations), while the senior research director, in consultation with the task force leads, developed a larger strategic logic based on the commissioner feedback: layered cyber deterrence.

FIGURE 4: THE SOLARIUM EVENT
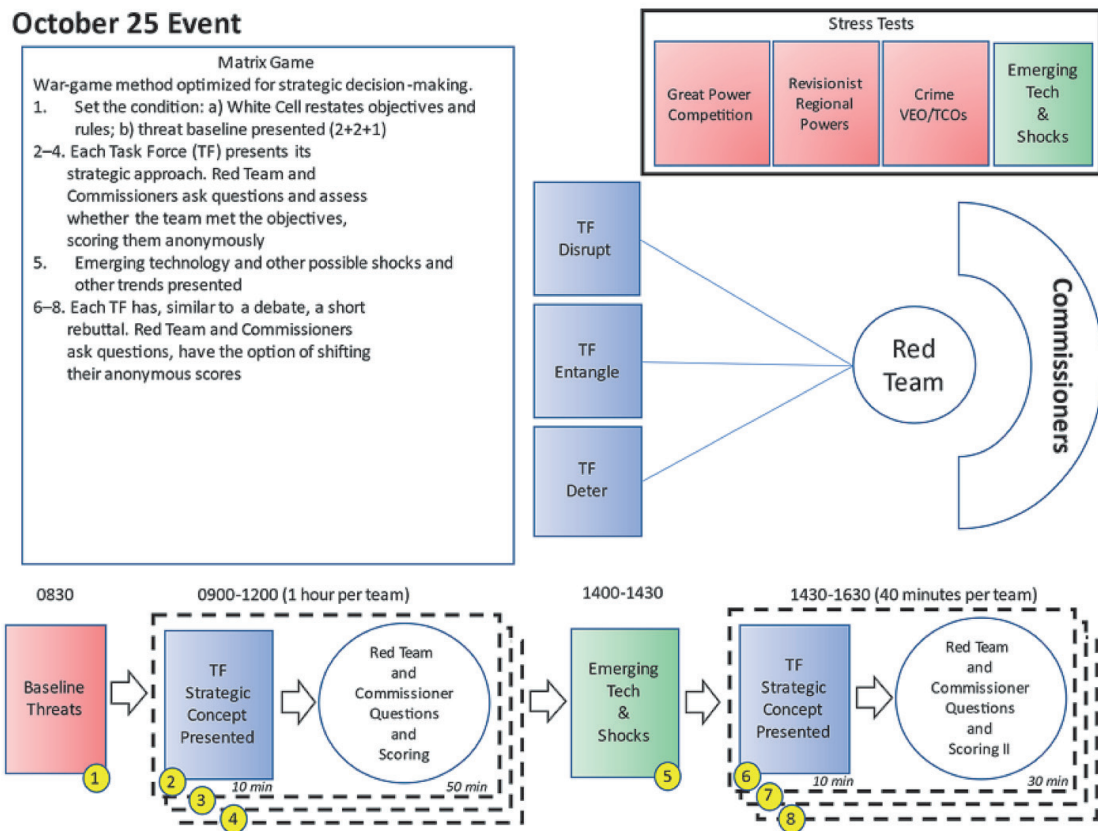
D. Layered Cyber Deterrence

The strategy of layered cyber deterrence emerged after months of work on an accelerated timeline. The Commission staff engaged hundreds of thought leaders, government officials, and stakeholders in the cyber security field. The outcome was a strategy encompassing three layers. Recognizing that we are in a period of neither war nor peace, layered cyber deterrence seeks to apply all levers of national power to the challenge of cyber conflict. The concept is consistent with the emergence of literature on competition in national security circles over the last 10 years, captured in the 2018 *Joint Concept for Integrated Campaigning* and the 2019 *Competition Continuum* (JD-19). The goal or victory condition of the strategy is to ensure that the connectivity required by modern society remains stable despite the cyber operations that target U.S. networks. Another goal is to reduce the severity of attacks below the threshold of armed conflict. Enabling this process depends on a new version of deterrence that moves us past the nuclear deterrence developed during the Cold War. Applying multiple instruments of power to ensure both survival and stability requires new ways to apply coercion in cyberspace.

Layered cyber deterrence relies on strong public-private collaboration to ensure that U.S. national cyber strategy does not remain siloed in the Department of Defense (DoD). The goal is to change the cost-benefit calculus of the adversary. The three layers provide overlapping visions of networked cyber strategy to protect the nation as it confronts new methods of digital warfare. The end state is to reduce the overall severity and frequency of cyber operations of significant consequence (Jensen 2020). The structure of the system and views of a persistent enemy do not dominate planning; rather, the strategy focuses on the interconnections built out through society that can enable security (Maoz 2010; Cleveland, Jensen et al. 2018).

The first layer is an outgrowth of entanglement strategies, meant to consider the global conditions critical for the development of cyber stability (Hurwitz 2012; Grigsby 2017; Nye 2017). The Solarium Commission seeks to create a plausible scenario that would enable allies to work together to create norms, institutions, and regulations that encourage responsible action in cyberspace. Excluding adversaries from this process creates bifurcated institutional systems that will hamper the development of cyber norms. Institutions can serve to facilitate agreements with allies and antagonists alike.

The development of norms is an often-theorized aspect of international relations, yet few have sought to understand what conditions create norms in the system (Finnemore and Hollis 2016). There has been a fruitful discussion of how to create legal norms (Schmitt and Vihul 2014), but engagement on the global institutional front is often stymied by geopolitical posturing (Grigsby 2017). Shaping the environment for action is critical, as digital connectivity depends on global networks and international collaboration to create a rules-based social order (Raymond 2021). In addition to negative aspects of coercion, the international environment can also enable positive methods of coercion that seek to change behavior through inducement rather than negative externalities (Baldwin 2020).

Recognizing that the United States cannot go on the offense until the home front is secure, the second layer advocates for denial strategies that ensure that the United States will be resilient in the face of inevitable cyber actions directed against the state (Gisladottir, Ganin et al. 2017; Valeriano and Jensen 2019). Deterrence by denial and target hardening will protect the networks from the most severe consequences of cyber actions (Denning 2014). Defense in being (mimicking the idea of a fleet in being) and defense in depth are both concepts that can be applied to cyber security (Hattendorf 2014). Only by enabling the defense can forward action take place, because then the homefront is not held at risk.

Finally, the third layer develops cost imposition as a strategy for the cyber domain. The imposition of costs is a critical method of applying force to coerce in cyberspace

and needed restoration after it was eliminated in persistent engagement. To prevent violations of critical thresholds and actions that seek to punish the civilian populations (Dev 2015), the United States must signal a strategy that will align consequences with deviant action. Signaling is a critical but often-forgotten aspect of international strategic positioning (Jervis 1970). In the layered cyber deterrence strategy, it became a critical mechanism enabling the means to achieve ends.

To enable cost imposition, the U.S. government will need to have its capabilities maintained and ready through proper force construction. The resource allocation dimensions of cyber security are often ignored, yet justifying how forces are arranged is critical in balancing the offense with the defense. There is also the possibility and potential for application of reserve forces in cyberspace to get beyond resource constraints (Hannan 2015). This became Recommendation 6.1.7, housed in the DoD given the current constraints of the Cybersecurity and Infrastructure Security Agency (CISA). The United States must ensure the cyber workforce can handle both the job of defensive and offensive action in and through cyberspace to harden targets and apply costs when needed.

Layered cyber deterrence does not seek to create a new paradigm for cyber security; rather, the strategy itself seeks to correctly apply a connection between the means and the ends to achieve clear victory conditions in cyber security. Rovner (2020) wonders critically just what the Solarium Commission rejects. The answer is limited strategies that seek to engage singular departments (USCYBERCOM and the DoD) that fail to conceive of a means to an end in strategy. These are explicitly rejected in favor of an end state that seeks overall U.S. stability and the reduction of attacks of consequence that hold the U.S. population at risk. We now move to evaluating the implementation of strategy, which is just as critical as the logical underpinnings of a strategy.

3. LEGISLATIVE SUCCESSES AND NATIONAL CYBER STRATEGY

A. The Legislative Strategy

The key task of the Solarium Commission was to enable reform. Commissions can be powerful venues for national security change (Tama 2011). As the Solarium Commission Report notes, "While cyberspace has transformed the American economy and society, the government has not kept up, and existing government structures limit cyber policymaking processes, dampen government action, and impede cyber operations" (Montgomery, Jensen et al. 2020, 2). Enabling success is critical, and it must come through action, not reports that tend to cycle throughout the U.S. government system. Almost every decade included a comprehensive evaluation of

cyber strategy with a series of recommendations for action, including the Ware Report (1970), the Cyberspace Policy Review (2009), and the Cyber Moonshot initiative (2018).

Overall, the Solarium Commission prioritized two tracks of effort to move past the failures of past efforts. One task was to build out and support the national cyber strategy of layered cyber deterrence. The second, and perhaps more important, task was to translate the specific legislative recommendations in the Solarium Report into law. In all, 52 legislative proposals made it into the report in Appendix B.[5] These proposals were extensively researched, supported by legislative analysis, and then distributed to the various committees and subcommittees in Congress. The goal was to find natural bipartisan support for each proposal as they became fully fleshed out and implementable law or directives to be included in legislation.

The Solarium Commission was able to get 25 of the 52 legislative recommendations written into the Fiscal Year 2021 National Defense Authorization Act (FY21 NDAA). In the end, 27 recommendations (two proposals were split) became law on January 1, 2021, after a veto override. Once the FY21 NDAA was signed into federal law, it became evident that the cyber security provisions included in the overall NDAA represent "the most comprehensive and forward-looking pieces of national cybersecurity in the nation's history."[6]

Some legislative recommendations failed to become law because either there was not enough time to develop the recommendations into full legislative proposals or there was no natural sponsor of legislation. The FY21 NDAA (Section 1714) authorized the Solarium Commission to continue its work for one more year to push through some recommendations that seek to improve cyber expertise in government (workforce), increase institutional cyber engagement (support the State Department), and enhance cyber reliance (in particular, to create a cyber recovery fund and develop breach notification law).

B. Evaluating the Legislative Successes

The two main successful legislative efforts sought either to enhance the power of existing cyber entities in the U.S. government or to create new structures to support the generation of a strategy to maintain security in cyberspace. While there is a need for a new cabinet-level organization to manage cyber security and information/data across the U.S. government, there is not much initiative to create such an organization due to the problems that developed after the creation of the DHS including complications at the border (Birkland 2009).

---

[5] More than 150 proposals were considered for the report; many of these were eliminated after the wargame.
[6] Statement by Solarium co-chairs Senator Angus King (I-Maine) and Representative Mike Gallagher (R-Wisconsin). https://www.king.senate.gov/newsroom/press-releases/solarium-co-chairs-welcome-26-recommendations-in-2021-national-defense-authorization-act.

The Commission therefore focused on enabling the functions of the U.S. government that could support cyber security efforts, with a focus on CISA, DoD, and USCYBERCOM. For example, Solarium Commission Recommendations 6.1 and 6.1.3 direct the DoD to conduct a force structure assessment of the Cyber Mission Force to ensure that USCYBERCOM has the resources needed to conduct operations that seek to impose costs (Section 1706 in the FY21 NDAA). The Solarium Commission also proposed that the DoD conduct an evaluation for the requirements needed to establish a cyber reserve force (Section 1730 and Recommendation 6.1.7) to support cyber mission forces.

To enable defensive operations, the Solarium Commission recommended vulnerability assessments to command-and-control functions of the DoD, including nuclear and conventional weapons systems (CSC Recommendation 6.2b and FY21 NDAA Section 1712). Another recommendation (CSC Recommendation 6.2.1 and 6.2.2) supported the need for the Defense Industrial Base (DIB) to participate in threat-intelligence-sharing programs (Section 1737) and threat-hunting on US networks (Section 1739). The Commission also enabled CISA to conduct threat-hunting investigations on US networks (Section 1705 of the FY21 NDAA and CSC Recommendation 1.4) and granted subpoena power to the organization (Section 1716 of the FY21 NDAA and CSC Recommendation 5.1.3).

C. National Cyber Director

Perhaps most importantly, the Commission recommended the creation of the position of a National Cyber Director (NCD) (hereafter, Recommendation 1.3), which became Section 1752 in the FY21 NDAA. The NCD position is meant to restore and to elevate a coordinator for all U.S. government efforts to establish a coherent whole-of-nation strategy for cyber security and to marshal incident response for major cyber breaches. Vesting such a position outside of the DoD and National Security Council, the NCD allows for the freedom of action to coordinate all sources of U.S. power towards the cyber domain, including the Department of Justice (indictments), the State Department (cyber diplomacy), and DHS (internal resilience).[7]

The Senate-confirmed position reporting directly to the president demonstrates the importance of the NCD coordinator position. Without such an office, the organizational seams (Chaudhary, Jordan et al. 2018) evident in the U.S. government will only continue to proliferate, endowing a disparate and uncoordinated cyber capability. Tasked with developing the overall U.S. cyber strategy, the NCD can help broaden how the U.S. considers cyber security as more than the domain of the U.S. military. Coordinating defensive efforts to respond to and survive a major cyber action highlights the importance of the position. The strategy of layered cyber deterrence

---

[7]   An National Security Council-housed cyber coordinator has limited ability to organize government responses and mainly focuses on ongoing threats, not the development of strategy and defenses to avoid attacks in the first place.

could become problematic if the layers end up working at cross purposes with each other. For example, the State Department's efforts to create viable norms can conflict with DoD offensive cyber impulses. Yet having a powerful NCD who can deconflict these issues and streamline processes is a critical task of this new role.

Even as some reject the need for reorganization of government (Rovner 2020), the Commission sought to focus on this key challenge to reform national strategy and process, preferring to not let bureaucratic divisions impede effective strategy. There was intense pushback on the NCD position from the Trump administration, because the bureaucratic power centers that developed during the administration were vested in those who sought to eliminate the White House cyber coordinator role in the first place. While the Biden administration has its own concerns about the NCD position, the main issue at this point is funding the organization and staff required to maintain a NCD position.

Finally, cyber security is a whole-of-nation challenge, not a whole-of-government problem. Most cyber resources, capabilities, and targets all reside beyond the control of the U.S. government. The NCD would be the point of contact for all private sector cyber stakeholders, ensuring there was an office that would be receptive to the needs of the private sector. In order to implement a national strategy, there needs to be one office that is responsible for coordination and strategic development that thinks beyond the bureaucratic demands of the specific cabinet-level branches.

4. PRESSURE POINTS AND MANAGING RISK

A. Cost Imposition and Enabling Defend Forward

Two early criticisms of the strategy of layered cyber deterrence are that it improperly returns the U.S. back to a deterrence strategy and that it revives the notion of the need to impose costs on the adversary. Persistent engagement is purposely framed as a natural evolution away from deterrence (Fischerkeller and Harknett 2017). Yet it is difficult to discard the concept of deterrence, given the demands of the policy community and a near-reflexive dependence on deterrence. The policy community tethers itself to deterrence as a process it knows and understands; there is a clear belief that nuclear deterrence has maintained stability during and after the Cold War.

The concept of layered deterrence is not about binary outcomes (cyber attack/no cyber attack). Rather, it is the mechanism to alter how states compete in cyberspace and the cascading effects cyber actions can have on global commerce given the dependence on connectivity. Layered cyber deterrence is a framework for competition more than it is a carbon copy of first-wave nuclear deterrence theory (Jervis 1978). Following the

original Solarium Commission model – not discarding it, as Rovner (2020) incorrectly charges – is a highlight of the deliberative process the Solarium Commission built to achieve consensus on cyber strategy.

In the cyber domain, there is a need to move past conventional notions of deterrence and rebuild the concept around the frames that are likely to enable cyber stability. Deterrence as articulated in the nuclear domain is the theory of preventing an action from happening through the threat of retaliation enabled by the ability to survive a first strike (Jervis 1978). Under this concept, cyber deterrence will never work because of the near constant probes and espionage attacks witnessed in cyberspace. Deterring cyber espionage, just like conventional espionage, is nearly impossible and too costly in relation to the benefits.

The goal instead is to reduce the severity and frequency of cyber activities. A state will never stop spying; what the target can do is make it harder for adversaries to spy on them, altering the expected value of the information they steal, and taking actions in the shadows that cause them to reconsider the logic of consequence associated with covert operations. This idea builds on new literature that finds that states use covert action to signal (Yarhi-Milo 2014; Carson and Yarhi-Milo 2017; Yarhi-Milo, Kertzer et al. 2018; Carson 2020). Layered cyber deterrence should therefore alter how states compete and deter attacks in the cyber domain above and below the threshold of armed conflict, including any provocative or disruptive actions that will inhibit the maintenance of information and command coordination capabilities. This can be done by creating the conditions in the system for the stable expectation of norms (shaping entanglement), denying attack surfaces to the opposition and enabling resilience in defense (denial), and by making clear, credible commitments to leverage consequences for deviant action (imposing costs).

As Fischerkeller and Harknett (2020) have noted in the past, "cost imposition is best understood as an effect resulting from the casual mechanism associated directly with a strategy of persistent engagement." In the hope of moving beyond coercion, persistent engagement discards cost imposition as a casual mechanism. A previous work of ours (Valeriano, Jensen et al. 2018) has suggested that coercion does not work in cyber competition; this finding has often been cited as evidence for the inability of coercion to achieve effects in cyberspace. That interpretation misunderstands the point of our work; it is not that coercion is impossible in cyberspace, but it is unlikely (Borghard and Lonergan 2017). This is often because the side that imposes costs does not clearly signal costs and has no credible commitment to follow through. Cyber operations are also better thought of as having a complementary and additive effect (Valeriano and Jensen 2021). Prior work demonstrates, when combining cyber operations data with event data on instruments of power, that all successful episodes of cyber coercion

occurred alongside a broader range of diplomatic, military, and economic inducements and threats (Valeriano, Jensen et al. 2018). Cyber operations are the icing, not the cake.

Persistent engagement had no clear identified causal mechanism connecting the ends and means because there is no clear end state. In failing to understand that the imposition of costs was not an outcome, but a feature of deterrence, persistent engagement has significant limits as a theory because it does not have a method of applying force against the adversary beyond friction (Fischerkeller and Harknett 2020). Without the imposition of costs, there is no conception of how to achieve an end (strategic stability through counter cyber operations) through a means (hunting forward). Friction is a useful method to confuse the adversary and distract their operations, but it is not a clear means to achieve an end because it depends on second- and third-order effects. The imposition of costs (along with resilience and entanglement) is the key element that makes the strategy of layered cyber deterrence effective. The remaining challenge is how to measure effectiveness and avoid escalation.

B. The Danger of Cyber Escalation

The prime risk associated with cyber security is the danger of a major cyber war that might destroy the economy, harm civilians, and disrupt critical infrastructure (all exaggerated fears but fears nonetheless) (Clarke and Knake 2014). This is a classic example of a low-probability, high-consequence risk, which, consistent with work on complex systems, could quickly evolve from a limited event to a systemic crisis. These dramatic actions would occur only after the confrontation between the entities engaged in serial competition escalates into violence. Understanding what escalation is and minimizing the risk of increasing intensity in cyber conflict was a task the Solarium Commission was not able to address through legislative recommendations, although it did study ways to minimize the risk. While layered cyber deterrence, if implemented, should stabilize cyber competition, there is still a systemic risk left to be addressed by future cohorts of academics, policy-makers, and activists.

The modern study of crisis escalation emerges during the Cold War through studies examining the process of bargaining during a foreign policy crisis (Schelling 1960; Schelling 1966). Kahn (1968) is the exemplar in the study of escalation, with his view that escalation results when one side tries to demonstrate resolve by increasing directed efforts in the diplomatic, military, information, or economic domains.

Escalation is defined as an increase in the intensity of conflict (vertical escalation) or to spread of the conflict to new venues (horizontal escalation). To escalate, Actor B (the target) must react with increased intensity after Actor A makes the first move. In cyberspace, this entails either reacting with more costly means of response using cyber options or by leveraging conventional operations to punish the initial

violation (Borghard and Lonergan 2019). Cyber escalation is an interactive process of increasing hostility and intensity over a series of interactions that occur in cyberspace. Libicki focuses on two factors: increasing the intensity of cyber operations (deeper, longer lasting effects) or finding more extensive cyber response options (striking new targets) (Libicki 2016).

Borghard and Lonergan (2019) argue that there is little logic behind the idea that cyber operations will provoke escalatory reactions, primarily because of the limited nature of the weapons, the uncertain effects, and the lack of costs imposed by cyber operations, meaning that the target often does not have to respond. Valeriano et al. (Valeriano, Jensen et al. 2018; Valeriano and Jensen 2021) go further by pointing out that cyber operations are ambiguous signals, used mostly as tools of espionage, that offer limited methods of coercion. Cyber operations can actually provide de-escalation pathways if utilized during a crisis to substitute for conventional operations (Valeriano and Jensen 2021).

Overall, the community has no clear idea about escalation patterns in cyberspace at this point because there is a limited availability of interactive data between adversaries. There is no data, as of yet, to establish a baseline of operations to understand how often operations fall above normal levels and demonstrate an increase in intensity. Empirically, there is evidence that escalation is rare in cyberspace, but these findings are based on data between rival actors (Valeriano, Jensen et al. 2018; Valeriano and Jensen 2019), wargames (Jensen and Banks 2018; Jensen and Valeriano 2019; Kreps and Schneider 2019), and surveys (Jensen and Valeriano 2019).

C. Managing the Risk of Cyber Escalation

Given the uncertainty we have on the probability of cyber escalation and what conditions provoke cyber dilemmas, it would be unwarranted to dismiss the possibility of escalation in the cyber domain. Thinking that offensive operations will not provoke retaliation seems to be prudent based on the evidence, but this evidence is limited.

The Obama administration era view of cyber strategy was focused on restraint to avoid "unintended damage and uncontrollable escalation" (Fischerkeller and Harknett 2017, 389). Observing that escalation is rare in the cyber domain – counting only two such incidents but without identifying the corpus of data – Fischerkeller and Harknett (2019) argue that states will establish a method of interaction based on agreed competition and avoid escalation.

Following this logic, some current U.S. cyber strategists seem to dismiss escalation concerns. Representatives of USCYBERCOM recently wrote: "Cyber Command takes these concerns seriously, and reducing the risk is a critical part of the planning

process. We are confident that this more proactive approach (persistent engagement) enables Cyber Command to conduct operations that impose costs while responsibly managing escalation" (Nakasone and Sulmeyer 2020). Confidence in managing the possibility of escalation does little to allay concerns that there will be escalation in the cyber domain due to provocative actions leveraged against an adversary.

The challenge is that managing escalation requires awareness of the dangers of escalation, clarity of national strategy, ability to signal intent to the opposition, data to observe risks, and institutions built to create a collaborative environment for problem solving. Therefore, the Solarium Commission submitted Recommendation 1.1.1, "Develop a Multitiered Signaling Strategy." The Commission Report notes, "Rather, the United States must signal capability and resolve, as well as communicate how it seeks to change adversary behavior and shape the strategic environment. Signaling is essential for escalation management so that actions taken in support of defend forward are not unintentionally perceived as escalatory" (Montgomery, Jensen et al. 2020, 33).

The signaling strategy should contain not only overt means of communication, including leveraging public diplomacy efforts and establishing clarity in national strategy, but also covert communications that seek to make clear the costs of deviant action in cyberspace. Proper communication is key to avoiding cyber disasters. No policy on signaling U.S. strategy was adopted by legislative recommendation, but a key task of the NCD (Section 1752) is to provide strategic leadership in cyber security, including coherently signaling cyber policy.

There is also a need to gather information and data on offensive cyber interactions to understand how these operations are received by the opposition. We know little about perceptions of U.S. action by adversaries. Do they understand U.S. strategy? Are there clear red lines in their estimation that forestall escalation? More intelligence would support better estimates of adversary perceptions. A breach notification law (Recommendation 4.7.1) would enable the collection of data on attacks on U.S. targets, helping strategies determine the impact of our operations on changing the behavior of the adversary.

Fostering more wargames in the cyber security community might help us understand the process of escalation better. This leads to Solarium Recommendation 3.3.4, which was the expansion of coordinated cyber exercises, gaming, and simulations. The FY2021 NDAA contains Section 1744, which establishes a biennial National Cyber Exercise. The goal of exercises is not to understand adversary reactions to U.S. strategy but to develop U.S. government agencies, private stakeholders, and international partners' experiences and processes when dealing with cyber threats. There needs to be a better concept of what metrics would be useful in establishing

the effectiveness strategy as it is implemented. Right now, we are flying blind and moving guideposts at will with no conception of benchmarks or methods to establish baselines.

5. THE CHALLENGE OF SOLARWINDS

A. What Was SolarWinds?

When the Solarium Commission tested its cyber strategies with a wargame, it developed two scenarios. Scenario 1 was *Slow Burn*, where a series of minor actions built up to create a crisis that demanded action from all U.S. government operations. The SolarWinds hack (Sanger, Perlroth et al. 2020) is exactly the sort of massive cyber operation that the Commission envisioned.

The SolarWinds operation targeted IT management software called Orion operated by the company SolarWinds. A supply-side vulnerability was exploited to insert malicious code that enabled hacker groups the Russian SVR or APT29 Cozy Bear (Sanger, Perlroth et al. 2020) to maintain a presence on U.S. networks and extract information at will. The complete fallout of the operation is still unknown.

The SolarWinds operation represents the future of digital political warfare, where rival states employ cyber operations to conduct limited operations meant to degrade or disrupt the capabilities of the opposition (Valeriano, Jensen et al. 2018). As a weak form of coercion, the espionage operation highlights the weaknesses in both the defenses and offensive capabilities of the United States as it operates in cyberspace.

B. The Failure of Persistent Engagement?

Some suggest the response to SolarWinds should include more persistent engagement operations. Harknett (2020), one of the original authors of the persistent engagement strategy (Fischerkeller and Harknett 2017), notes that "the United States must accelerate its adoption of the doctrine of persistent engagement across the entirety of its intergovernmental space…. Had the doctrine been in place fully and comprehensively, the form of this attack and its consequences may have been different."

Harknett (2020) notes that the USCYBERCOM mission set is limited to protecting the Defense Information Network. As Corn (2021) notes, "as for allegations that Cyber Command failed to defend forward in this instance, the charge presumes without public evidence that, among other things, the Defense Department and Cyber Command were provisioned with the authority to disrupt SolarWinds." By implication, the suggestion is that USCYBERCOM needs to implement more defend-forward operations and needs more legal authorities to do so to fulfill its mission.

If the U.S. loses the initiative, Russia might dictate the pace of cyber operations and place a constant stress on U.S. defense, which would lead to U.S. failure, according to Harknett (2020). Instead, the SolarWinds operation highlights the limitations of persistent engagement as the operationalization of defend forward (Nakasone 2019). There is a clear role for defend forward operations in cyberspace, but as the sole form of forward operations, said strategies can be self-defeating, because we lack a conception of how the opposition will receive such operations. In fact, they will likely provoke counter and proportional operations that use the same strategy against the defender, which might be exactly how the Russians conceive of the SolarWinds operation. A poorly signaled strategy may well encourage them precisely to counter defend forward operations with their own forward operations.

Persistent engagement lacks a strategy of imposing clearly signaled costs on the opposition, so the opposition has freedom of movement. National strategy needs to be clarified to impose costs and create normative/legal restraints for violations like SolarWinds. Forward maneuver doctrines can only be sustained with strong defenses and a clear strategy of imposing costs on the adversary for deviant actions.

C. The Failure of the Defense?

There is also the need to truly conceptualize what defend forward means in operation. As Borghard and Schneider (2020) note, "we see [defend forward] as two types of activities: The first is information gathering and sharing with allies, partner agencies, and critical infrastructure by maneuvering in networks where they operate." By establishing more entangling partnerships in the international system and facilitating more cooperation with the private sector (Raymond and DeNardis 2015), the U.S. government should be better able to enable the protection of its networks through information-sharing. Forward operations require not only threat-hunting but also creating the overall conditions conducive to denial operations.

In the future, a deeper focus on denial-based strategies outlined in Layer 2, "deny benefits," is critical. Enabling CISA to launch internal threat-hunting would foster an environment for innovation where the continuous monitoring systems could be updated to be more proactive against unknown threats. Utilizing subpoena authority now granted to CISA, the U.S. government can more effectively implement defensive operations.

Making espionage activities more costly and difficult is the goal. The attacker is then limited in their options and must expend added effort to succeed, which thereby decreases the severity and frequency of attacks. By focusing on more than the offense, under the coordination of the NCD, the U.S. can seek to implement a cyber strategy that carefully considers the utility of defensive operations alongside hunting forward.

6. PATH FORWARD

The Solarium Commission will likely endure as a singularly effective effort to construct a roadmap for national cyber strategy. By basing the Solarium Commission Report on research, evidence, and data, the Solarium Commission sought to develop a unique strategy that considers the offense, defense, and systemic constraints at the same time, moving beyond the monocausal strategies developed in the past.

The other key innovation was thinking of cyber strategy in an integrated-network sense. The Solarium Commission began by developing a whole-of-nation strategy that sought to include both public and private stakeholders in seeking to defend the nation. This pushes the cyber security community to think more about how network connectivity is both a strength and a weakness for society. In short, the entire nation needs to be involved in the effort of cyber security, because attack surfaces in the United States are so vast.

The Solarium Commission was successful in getting a majority of its recommendations enacted into law, putting a force behind the ideas it developed that seek to ensure that cyber strategy becomes a continual and evolving process. The U.S. needs to build on its successes and avoid developing a new strategy for every new administration. The Solarium Commission will continue its work for the rest of 2021 to support the Biden administration in implementing its recommendations. Hopefully, the next Commission or strategy review does not have to repeat the effort again in five years.

The development of strategy needs to move beyond the impulses of particular departments (like the DoD) or administrations, because bureaucratic political considerations can become the enemy of progress and fail to engage the marketplace of ideas. People and organizations fall in love with their ideas over time and fail to think about the evaluation of strategies, because they become doctrinal. Policy is often the art of compromise; the Solarium Commission process was as different as it was similar to the original Eisenhower Solarium effort, because it valued bipartisan compromise, academic research, community advice, and empirical verification. If anything, the process was more inclusive and academically rigorous, providing hope that the community can avoid repeating past arguments and debates.

What remains is how the achievements of the Solarium Commission, including the NCD position, will evolve over time. Other countries can take this process as a model for their own strategic reform or, possibly, a model to avoid if the U.S. continues to fall into the trap of the pathologies of the past (not enabling cost imposition, weak defenses, or not shaping the norms and regulations that guide the system). Only time will be the judge.

REFERENCES

Baldwin, D. A. 2020. *Economic Statecraft*. New edition. Princeton University Press.

Birkland, T. A. 2009. "Disasters, Catastrophes, and Policy Failure in the Homeland Security Era." *Review of Policy Research* 26, no. 4: 423–438. https://doi.org/10.1111/j.1541-1338.2009.00393.x.

Borghard, E. D., and S. W. Lonergan. 2017. "The Logic of Coercion in Cyberspace." *Security Studies* 26, no. 3: 452–481. https://doi.org/10.1080/09636412.2017.1306396.

Borghard, E. D., and S. W. Lonergan. 2019. "Cyber Operations as Imperfect Tools of Escalation." *Strategic Studies Quarterly* 13, no. 3: 122–145. https://www.jstor.org/stable/26760131.

Borghard, E. D., and J. Schneider. 2020. "Russia's Hack Wasn't Cyberwar. That Complicates US Strategy." *Wired*, December 17. Accessed April 24, 2021. https://www.wired.com/story/russia-solarwinds-hack-wasnt-cyberwar-us-strategy/.

Carson, A. (2020). *Secret Wars: Covert Conflict in International Politics*. Princeton University Press.

Carson, A., and K. Yarhi-Milo. 2017. "Covert Communication: The Intelligibility and Credibility of Signaling in Secret." *Security Studies* 26, no. 1: 124–156. https://doi.org/10.1080/09636412.2017.1243921.

Chaudhary, T., J. Jordan, M. Salomone, and P. Baxter. 2018. "Patchwork of Confusion: The Cybersecurity Coordination Problem." *Journal of Cybersecurity* 4, no. 1: 1-13. https://doi.org/10.1093/cybsec/tyy005.

Clarke, R. A., and R. K. Knake. 2014. *Cyber War*. Old Saybrook, CT: Tantor Media.

Cleveland, C., B. M. Jensen, A. David, and S. F. Bryant. 2018. *Military Strategy for the 21st Century: People, Connectivity, and Competition*. Cambria Press.

Cohen, M. D., J. G. March, and J. P. Olsen. 1972. "A Garbage Can Model of Organizational Choice." *Administrative Science Quarterly* 17, no. 1 (March): 1–25. https://doi.org/10.2307/2392088.

Congress. 2017–2018. H.R. 5515 – John S. McCain National Defense Authorization Act for Fiscal Year 2019. U. S. Congress. Washington, DC: U.S. Government Printing Office. https://www.congress.gov/115/plaws/publ232/PLAW-115publ232.pdf.

Corn, G. 2021. "SolarWinds is Bad, but Retreat From Defend Forward Would Be Worse." *Lawfare*, January 14. Accessed April 24, 2021. https://www.lawfareblog.com/solarwinds-bad-retreat-defend-forward-would-be-worse.

Denning, D. E. 2014. "Framework and Principles for Active Cyber Defense." *Computers and Security* 40: 108–113. https://doi.org/10.1016/j.cose.2013.11.004.

Dev, P. R. 2015. "Use of Force and Armed Attack Thresholds in Cyber Conflict: The Looming Definitional Gaps and the Growing Need for Formal UN Response." *Texas International Law Journal* 50: 381.

Durant, R. F., and P. F. Diehl. 1989. "Agendas, Alternatives, and Public Policy: Lessons from the US Foreign Policy Arena." *Journal of Public Policy* 9, no. 2: 179–205.

Finnemore, M., and D. B. Hollis. 2016. "Constructing Norms for Global Cybersecurity." *American Journal of International Law* 110, no. 3: 425–479.

Fischerkeller, M. P., and R. J. Harknett. 2017. "Deterrence is Not a Credible Strategy for Cyberspace." *Orbis* 61, no. 3: 381–393. https://doi.org/10.1016/j.orbis.2017.05.003.

Fischerkeller, M. P., and R. J. Harknett. 2019. "Persistent Engagement, Agreed Competition, and Cyberspace Interaction Dynamics and Escalation." *Cyber Defense Review* (special issue): 267–287. https://cyberdefensereview.army.mil/Portals/6/CDR-SE_S5-P3-Fischerkeller.pdf.

Fischerkeller, M. P., and R. J. Harknett. 2020. "Persistent Engagement and Cost Imposition: Distinguishing Between Cause and Effect." *Lawfare*, February 6. Accessed April 24, 2021. https://www.lawfareblog.com/persistent-engagement-and-cost-imposition-distinguishing-between-cause-and-effect.

Gallagher, M. J. 2015. "Intelligence and National Security Strategy: Reexamining Project Solarium." *Intelligence and National Security* 30, no. 4: 461–485. https://doi.org/10.1080/02684527.2014.885203.

Gisladottir, V., A. A. Ganin, J. M. Keisler, J. Kepner, and I. Linkov. 2017. "Resilience of Cyber Systems with Over and Underregulation." *Risk Analysis* 37, no. 9: 1644–1651. https://doi.org/10.1111/risa.12729.

Grigsby, A. 2017. "The End of Cyber Norms." *Survival* 59, no. 6: 109–122. https://doi.org/10.1080/00396338.2017.1399730.

Hannan, N. K. 2015. "Use of Reserve Forces in Support of Cyber-Resilience for Critical National Infrastructure: US and UK Approaches." *RUSI Journal* 160, no. 5: 46–51. https://doi.org/10.1080/03071847.2015.1102543.

Harknett, R. J. 2020. "SolarWinds: The Need for Persistent Engagement." *Lawfare*, December 23. Accessed April 24, 2021. https://www.lawfareblog.com/solarwinds-need-persistent-engagement.

Hattendorf, J. B. 2014. "The Idea of a 'Fleet in Being' in Historical Perspective." *Naval War College Review* 67, no. 1: 42–60. https://digital-commons.usnwc.edu/nwc-review/vol67/iss1/6/.

Healey, J. 2019. "The Implications of Persistent (and Permanent) Engagement in Cyberspace." *Journal of Cybersecurity* 5, no. 1: https://doi.org/10.1093/cybsec/tyz008.

Hurwitz, R. 2012. "Depleted Trust in the Cyber Commons." *Strategic Studies Quarterly* 6, no. 3: 20–45. https://www.jstor.org/stable/26267260.

Jensen, B. M. 2018. "The Role of Ideas in Defense Planning: Revisiting the Revolution in Military Affairs." *Defence Studies* 18, no. 3: 302–317. https://doi.org/10.1080/14702436.2018.1497928.

Jensen, B., and D. Banks. 2018. *Cyber Operations in Conflict: Lessons from Analytic Wargames*. Center for Long-Term Cybersecurity, UC Berkeley. https://cltc.berkeley.edu/2018/04/16/cyber-operations-conflict-lessons-analytic-wargames/.

Jensen, B., and B. Valeriano. 2019. *Cyber Escalation Dynamics: Results from War Game Experiments*. International Studies Association, Annual Meeting, Toronto, Ontario, Canada. http://web.isanet.org/Web/Conferences/Toronto%202019-s/Archive/71e7820c-e61c-4187-ab8c-28de83dfd660.pdf.

Jensen, B., and B. Valeriano. 2019. *What Do We Know about Cyber Escalation? Observations from Simulations and Surveys*. Atlantic Council. Accessed April 24, 2021. https://www.atlanticcouncil.org/in-depth-research-reports/issue-brief/what-do-we-know-about-cyber-escalation-observations-from-simulations-and-surveys/.

Jensen, B. 2020. "Layered Cyber Deterrence: A Strategy for Security Connectivity in the 21st Century." *Lawfare*, March 11. https://www.lawfareblog.com/layered-cyber-deterrence-strategy-securing-connectivity-21st-century.

Jervis, R. 1970. *The Logic of Images in International Relations*. Princeton, NJ: Princeton University Press.

Jervis, R. 1978. "Deterrence Theory Revisited." *World Politics* 31, no. 2: 289–324.

Kahn, H. 1968. *On Escalation: Metaphors and Scenarios*. Transaction Publishers.

Kingdon, J. W., and E. Stano. 1984. *Agendas, Alternatives, and Public Policies*. Boston: Little, Brown.

Klimburg, A. 2012. *National Cyber Security Framework Manual*. NATO Cooperative Cyber Defense Center of Excellence. https://www.ccdcoe.org/uploads/2018/10/NCSFM_0.pdf.

Kreps, S., and J. 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.

Libicki, M. 2016. *Cyberspace in Peace and War*. Naval Institute Press.

Maoz, Z. 2010. *Networks of Nations: The Evolution, Structure, and Impact of International Networks, 1816–2001*. Cambridge University Press.

Mintrom, M. 1997. "Policy Entrepreneurs and the Diffusion of Innovation." *American Journal of Political Science* 41, no. 3 (July): 738–770. https://doi.org/10.2307/2111674.

Montgomery, M., B. Jensen, E. D. Borghard, J. Costello, V. Cornfeld, C. Simpson, and B. Valeriano. 2020. *Cyberspace Solarium Commission Report*. Washington, DC. https://www.solarium.gov/report.

Nakasone, P. M. 2019. "A Cyber Force for Persistent Operations." *Joint Force Quarterly* 92: 10–14. http://cs.brown.edu/courses/csci1950-p/sources/2019_01_22_JFQ_CyberRoleForPersistentOperations_Nakasone.pdf.

Nakasone, P. M., and M. Sulmeyer. 2020. "How to Compete in Cyberspace." *Foreign Affairs*, August 25. https://www.foreignaffairs.com/articles/united-states/2020-08-25/cybersecurity.

Nye, J. S., Jr. 2017. "Deterrence and Dissuasion in Cyberspace." *International Security* 41, no. 3: 44–71. https://doi.org/10.1162/ISEC_a_00266.

Raymond, M. 2021. "Social Practices of Rule-Making for International Law in the Cyber Domain." *Journal of Global Security Studies* 6, no. 2: https://doi.org/10.1093/jogss/ogz065.

Raymond, M., and L. DeNardis. 2015. "Multistakeholderism: Anatomy of an Inchoate Global Institution." *International Theory* 7, no. 3: 572–616. https://doi.org/10.1017/S1752971915000081.

Rovner, J. 2020. "Did the Cyberspace Solarium Commission Live Up to Its Name?" *War on the Rocks*, March 19. Accessed April 24, 2021. https://warontherocks.com/2020/03/did-the-cyberspace-solarium-commission-live-up-to-its-name/.

Sanger, D. E., N. Perlroth, and E. Schmitt. 2020. "Scope of Russian Hacking Becomes Clear: Multiple U.S. Agencies Were Hit." *New York Times*, December 14. Accessed April 24, 2021. https://www.nytimes.com/2020/12/14/us/politics/russia-hack-nsa-homeland-security-pentagon.html.

Schelling, T. 1960. *The Strategy of Conflict*. Harvard University Press.

Schelling, T. C. 1966. *Arms and Influence*. New Haven: Yale University Press.

Schmitt, M. N., and L. Vihul. 2014. "The Nature of International Law Cyber Norms." *Tallinn Papers* (no. 5). https://ssrn.com/abstract=2543520.

Tama, J. 2011. *Terrorism and National Security Reform: How Commissions Can Drive Change During Crises*. Cambridge University Press.

Valeriano, B., B. M. Jensen, and R. C. Maness. 2018. *Cyber Strategy: The Evolving Character of Power and Coercion*. New York: Oxford University Press.

Valeriano, B. G., and B. Jensen. 2019. "The Myth of the Cyber Offense: The Case for Cyber Restraint." *Cato Institute Policy Analysis* (no. 862). https://www.cato.org/policy-analysis/myth-cyber-offense-case-restraint.

Valeriano, B., and B. Jensen. 2021. "De-Escalation Pathways and Disruptive Technology: Cyber Operations as Off-Ramps to War." In *Cyber Peace*, edited by S. Shackelford. Cambridge University Press.

Yarhi-Milo, K. 2014. *Knowing the Adversary: Leaders, Intelligence, and Assessment of Intentions in International Relations*. Princeton University Press.

Yarhi-Milo, K., J. D. Kertzer, and J. Renshon. 2018. "Tying Hands, Sinking Costs, and Leader Attributes." *Journal of Conflict Resolution* 62, no. 10: 2150–2179. https://doi.org/10.1177%2F0022002718785693.

Zenko, M. 2015. *Red Team: How to Succeed by Thinking Like the Enemy*. Basic Books.

The Cyberspace 'Great Game'. The Five Eyes, the Sino-Russian Bloc and the Growing Competition to Shape Global Cyberspace Norms

Nikola Pijović

Cyber Security Cooperative Research Centre Postdoctoral Research Fellow
School of Politics and International Relations / Law School
University of Adelaide, Australia
nikola.pijovic@adelaide.edu.au

Abstract: The development of global norms of responsible state behaviour in cyberspace has, over the past decade, become a significant foreign policy issue and a new battleground between states. The contested and competitive nature of global cyberspace norm building suggests that although there are complicated legal and technical issues at play, the development of cyberspace norms remains primarily a contestation of values, ideologies, and strategic interests. This paper argues that the competition to shape the governance of cyberspace through the development of cyberspace norms represents a continuation of foreign and strategic policy applied to the cyber domain. This has resulted in a growing cyberspace 'Great Game' between the Five Eyes alliance countries (the United States, United Kingdom, Australia, Canada, and New Zealand) and the Sino-Russian bloc (China and Russia). The Five Eyes and the Sino-Russian bloc are key cyber powers and cyberspace norm entrepreneurs whose leadership is instrumental in promoting global cyberspace norm preferences. However, each camp advocates a set of norm preferences inherently at odds with the other's, which has resulted in growing competition for dominance in cyberspace norm prescription and promotion. The paper outlines the key cyberspace norm proposals and initiatives promoted by the Five Eyes and the Sino-Russian bloc, discussing their main differences. It argues that the latest round (2019–2021) of the United Nations Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace (UNGGE) deliberations is unlikely to help bridge these differences in any substantive way. The cyberspace 'Great Game' and the increasingly competitive

nature of cyberspace norm development will remain a feature of global efforts to govern cyberspace throughout the 2020s.

Keywords: *Five Eyes, cyberspace norms, China, Russia, cyber security*

1. INTRODUCTION[1]

The governance of cyberspace and development of global norms of responsible state behaviour have, over the past decade, become significant international relations issues. Although the decade of the 2000s saw limited efforts aimed at international agreement over the governance of cyberspace, the 2010s have seen a proliferation of global cyberspace norms (Barrinha and Renard 2020, 764). However, the contested and competitive nature of global cyberspace norm building suggests that although there are complicated legal and technical issues at play, the development of cyberspace norms remains inherently a contestation of values, ideologies, and interests. Echoing Carl von Clausewitz's (2007, 28) famous maxim that war 'is merely the continuation of policy by other means', this paper argues that notwithstanding the novelty of the cyber domain, the development of cyberspace norms is merely a continuation of foreign and strategic policy by other means. This continuation of foreign and strategic policy has resulted in a cyberspace 'Great Game' between the Five Eyes alliance (the United States [US], United Kingdom [UK], Australia, Canada, and New Zealand) and the Sino-Russian bloc (China and Russia). The Five Eyes and the Sino-Russian bloc compete for dominance in cyberspace norm prescription, with each side advocating a set of norm preferences incompatible with the other's. While these norm preferences are also advocated by, and find varying degrees of support in, many other states, the Five Eyes and the Sino-Russian bloc merit special attention as key global cyber powers and norm entrepreneurs whose leadership is instrumental in promoting and adhering to global cyberspace norm preferences.[2]

The paper first provides brief background on the Five Eyes and the Sino-Russian bloc as global cyber powers. It then examines their competing cyberspace norm preferences and norm building initiatives. The focus is on critically analysing the fundamental differences underpinning each side's differing conceptions of what global cyberspace norms should promote: 'cyber security' versus 'information security', 'multi-

---

[2]   These states are not the only globally relevant cyber powers, but they rank among the top 10 most cyber powerful nations and have the most advanced abilities to 'conduct aggressive cyber operations (or to deter or withstand such operations)', 'influence the international cyber agenda', and use cyber tools to 'promote a broader agenda and wider interests' (Barrinha and Renard 2020, 755). For cyber power rankings, see Voo et al. (2020), 8.

stakeholder' versus 'multilateral' governance of the internet, and 'transnationalism' versus 'cyber sovereignty'. The conclusion highlights the implications of the cyberspace 'Great Game' for the future of cyberspace norm development.

2. THE FIVE EYES AND THE SINO-RUSSIAN BLOC AS KEY CYBERSPACE POWERS

In the original 'Great Game', the British and the Russians competed for influence in Central Asia throughout the 19th century. In the cyberspace 'Great Game', the British and the Russians, in addition to other key cyber powers such as the US and China, compete over how the world governs cyberspace. As in the original 'Great Game', there is no overt military confrontation (yet), but unlike the original 'Great Game', the fallout of the cyberspace one is truly global in reach. For example, as the Five Eyes countries have moved to effectively ban the Chinese company Huawei from participating in the construction of their 5G mobile technology infrastructure (Slezak and Bogle 2018; Trump 2019; Tobin 2019; Gold 2020; Duckett 2020), international economists have warned that the ban poses a significant threat to the stability and growth of the global economy (Moon and Bray 2019). Although this ban has been discussed primarily in terms of national security (could the Chinese government use Huawei's technology for spying purposes?), it is fundamentally underpinned by the contestation of ideas about the governance of cyberspace and norms of responsible state behaviour. Given that 5G technology is the future of cyberspace and global connectivity and the world is – especially as a result of the COVID-19 pandemic – increasingly dependent on information and communications technology (ICT), the Huawei story aptly highlights the global significance of the cyberspace 'Great Game'. Therefore, while there are many differences between the original 'Great Game' and the one now played in cyberspace, the comparison points out the highly competitive nature of the global governance of cyberspace.

The Five Eyes are made up of the US, UK, Australia, Canada, and New Zealand, united 'by the language, values and institutions associated with the historical experience' of Britain's empire (Vucetic 2010, 456). This grouping is also referred to as the 'Anglosphere' (Wellings and Mycock 2019). Although not a 'unitary actor' in global affairs, the Anglosphere continues to 'define, order and promote' the values, policies, and transnational institutions that underpin the current rules-based international order (Vucetic 2010, 469). This is mainly due to the Five Eyes' success in fashioning the post-World War II global order, whereby 'a global society hitherto dominated by a system of states and empires received an important layer of multilateral institutions designed mostly by, and for, the Anglo-American elites' (Vucetic 2010, 468).

However, the close relations between the Five Eyes are today underpinned by more than just a shared history and ability to shape global power dynamics. They are underpinned by a closely aligned strategic interest, especially vis-à-vis China, and an ever-growing web of Five Eyes 'policy networks' that 'have been central to the co-production of policy, collaboration in shared policy problems and the transfer of policy ideas and practices' between these countries (Legrand 2019, 66). The Five Eyes alliance itself was established by the 1946 British-US Communication Intelligence Agreement (UKUSA), updated in 1955 and expanded to include Australia, Canada, and New Zealand (NSACSS n.d.).[3] Today, the Five Eyes constitute 'a cooperative, complex network of linked autonomous intelligence agencies' where 'individual intelligence organizations follow their own nationally legislated mandates, but interact with an affinity strengthened by their common Anglo-Saxon culture, accepted liberal democratic values and complementary national interests, all seasoned with a profound sense of confidence in each other and a degree of professional trust so strong as to be unique in the world' (Cox 2012).

Cyber security is of critical importance for the Five Eyes alliance, with the original UKUSA Agreement founded on the sharing of 'communications intelligence' – today's 'signals intelligence' (SIGINT). The continued centrality of SIGINT sharing to the Five Eyes alliance is reflected in the fact that the core Five Eyes intelligence agencies remain SIGINT and cryptology agencies: the National Security Agency (US), the Government Communications Headquarters (UK), the Australian Signals Directorate (AUS), the Communications Security Establishment (CAN), and the Government Communications Security Bureau (NZ) (Richelson 2016, 370). The Five Eyes countries enjoy some of the highest internet usage rates globally; their economies are increasingly digital and therefore significantly exposed to the benefits and perils of cyberspace (especially in the COVID-19 pandemic environment); and their national security infrastructure and governance systems are overwhelmingly reliant on ICT – all clear reasons why cyber security holds critical importance (GCDL n.d.).

China and Russia also rank among the world's strongest cyber powers. In the past two decades they have developed a clear strategic closeness, largely motivated by concerns over the West's (and especially the Five Eyes') promotion of global political and economic liberalism in the post-Cold War period (Bolt and Cross 2018; Lukin 2018). This unity of purpose in contesting many of the values, policies, and norms associated with global liberalism is especially evident with regards to the governance of cyberspace. China's cyber strategy is underpinned by a fundamental focus on cyber sovereignty, exhibited in three key goals: limiting the threat of the internet to the Communist Party's hold on power; shaping global cyberspace norms to extend China's political, military, and economic influence; and countering Five Eyes advantages in cyberspace (Segal 2017, 1). While appreciation of cyber sovereignty is, to varying

---

3   The actual term 'Five Eyes' refers to a dissemination caveat of intelligence products ('Secret/Top Secret – AUS/CAN/NZ/UK/US Eyes Only') shortened by practitioners in everyday use. See Cox (2012).

degrees, shared by all states (including the Five Eyes), the primacy of sovereignty in China's 'cyber diplomacy' has become a significant point of contestation with the Five Eyes, who advocate for a more open and transnational cyberspace. In the past decade, China's cyber security policy has consistently highlighted sovereignty as a key concern, with its 2016 *National Cyberspace Security Strategy* (CCM 2016) and 2017 *International Strategy for Cooperation in Cyberspace* (MFAPRC 2017) ranking sovereignty as a first principle and strategic objective.

Russia's cyber strategy is underpinned by the concept of 'information warfare', with a preference for the term 'information security' (a preference shared by China). The Russians define information warfare as '…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', constituting 'a massive psychological manipulation of the population to destabilize the state and society' and compelling 'the state to take decisions for the benefit of the opposing force' (Lilly and Cheravitch 2020, 133). Russia's strategic thinking on information/cyber security continues to be dominated by the idea of information warfare and supremacy, with a particular concern about foreign interference (Lilly and Cheravitch 2020, 134–136). These concerns are generally shared by China and motivate both countries to promote global cyberspace norms they hope will advance their interests and constrain the Five Eyes' cyber dominance. While there are many differences between China and Russia's cyber security experiences (Whyte and Mazanec 2019, 232), their primary concern over political stability and foreign interference is a key factor underpinning their cooperation on the development of global cyberspace norms.

3. HOW THE FIVE EYES AND THE SINO-RUSSIAN BLOC ARE SHAPING THE GOVERNANCE OF CYBERSPACE

Since the early years of the 21st century, the Five Eyes and the Sino-Russian bloc have played prominent roles in trying to shape how cyberspace is governed. Although they have been able to nominally agree that existing international law applies to the cyber domain, the two blocs hold inherently incompatible cyberspace norm preferences. Their contestation of ideas on how cyberspace should be governed is underpinned by fundamental ideological differences in conceptualizing cyber security and the political values each bloc promotes in an effort to advance their strategic interests. Before examining their differing visions for the governance of cyberspace, it is worth briefly reiterating what the two blocs have been able to at least nominally agree on. In this context, the paper examines their contributions to the most prominent global forum discussing the governance of cyberspace: the United Nations Group

of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace (formerly, the United Nations Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security) (UNGGE).[4]

There have been six UNGGE sessions: 2004–2005, 2009–2010, 2012–2013, 2014–2015, 2016–2017, and 2019–2021 (currently still underway). Reports are produced by consensus, with the whole group having to 'agree upon the report in its entirety' before making it public, and this implies at least nominal (if not always substantive) agreement (UNIDIR 2016, 5). The first UNGGE (2004–2005) failed to reach a consensus report, with differences arising over 'how to characterize the threat posed by State exploitation of ICTs for military purposes' and 'whether the discussion of ICT security should focus solely on the ICT infrastructure or include information content as well' (UNIDIR 2016, 6). The second UNGGE (2009–2010) produced a report which expressed its concern about states 'developing ICTs as instruments of warfare and intelligence, and for political purposes', recommending 'further dialogue among States to discuss norms pertaining to State use of ICTs, to reduce collective risk and protect critical national and international infrastructure' (UN A/Res/65/201 [2010], 2, 8). The third UNGGE (2012–2013) agreed that 'international law, and in particular the Charter of the United Nations, is applicable and is essential to maintaining peace and stability and promoting an open, secure, peaceful and accessible ICT environment'. It also concluded that state sovereignty applies in cyberspace, that 'state efforts to address the security of ICTs must go hand-in-hand with respect for human rights and fundamental freedoms set forth in the Universal Declaration of Human Rights and other international instruments', and that states 'must meet their international obligations regarding internationally wrongful acts attributable to them', 'must not use proxies to commit internationally wrongful acts', and 'should seek to ensure that their territories are not used by non-State actors for unlawful use of ICTs' (UN A/Res/68/98 [2013], 8). The fourth UNGGE (2014–2015) reiterated the same points made in the third session, adding that states had 'jurisdiction over the ICT infrastructure located within their territory' and that 'the accusations of organizing and implementing wrongful acts brought against States should be substantiated' (UN A/Res/70/174 [2015], 12, 13).

The fifth UNGGE (2016–2017) was to expand on the question of 'how' international law applied to norms of responsible state behaviour but was unable to find agreement. It failed to provide a consensus report for various political and ideological reasons, confirming 'that there are significant differences of opinion' between states 'on how to apply international law' to their use of ICTs, and that the most visible and sensitive contestation is related to questions of 'state sovereignty versus international obligations, and the relationship between the State and the individual' (Tikk and Kerttunen 2017, 15). The divisions were familiar: between the 'Western or "like-

---

4   China, Russia, the UK, and the US are permanent members of the UNGGE. Australia was a member in 2012–2013, 2016–2017, and 2019–2021, and Canada in 2012–2013, 2014–2015, and 2016–2017.

minded" approach that focuses on promoting and explaining' existing international law's applicability to cyberspace, and the Sino-Russian 'call for *lex specialis*' to govern cyberspace and 'reinforced international political structures, mainly the UN, as the mechanism to maintain international peace and security' (Tikk and Kerttunen 2017, 15–16). As of the writing of this paper, it is unclear if the sixth UNGGE (2019–2021) will produce a consensus report.

To sum up, the UNGGE process has been able to establish at least some basic principles on the applicability of international law to cyberspace, which can serve as a starting point for global cyberspace norm building. However, these broad categories of international law – often subject to differing and competing interpretations – allow both the Five Eyes and the Sino-Russian bloc to claim legitimacy for their respective cyberspace norm preferences as embedded in international law.

A. The Five Eyes' cyber security norm preferences

The Five Eyes have collectively published 16 national cyber security strategy documents,[5] and although only the ones published in the past decade focus substantively on cyberspace norms, this body of documents clearly indicates Five Eyes cyberspace norm preferences. The May 2011 US *International Strategy for Cyberspace* was the first Five Eyes cyber strategy to outline a clear position on norms of responsible state behaviour in cyberspace. The goal was to 'promote an open, interoperable, secure, and reliable information and communications infrastructure' by building an environment 'in which norms of responsible behaviour guide states' actions… and support the rule of law in cyberspace' (US 2011, 8). The key principles outlined included upholding fundamental freedoms like association and expression; respect for intellectual property and copyright; online privacy and the protection from arbitrary or unlawful state interference with citizens' use of the internet; protection from cyber crime; support for a multi-stakeholder management of the internet; and the right to self-defence by 'all necessary means' (which may be triggered by aggressive malicious acts in cyberspace) (US 2011, 10–14). These principles still form the core Five Eyes cyberspace norm preferences.

In November 2011, the British government published *The UK Cyber Security Strategy: Protecting and Promoting the UK in a Digital World*, discussing 'rules of the road' for state behaviour in cyberspace. The UK's position was that 'all governments must act proportionately in cyberspace and in accordance with national and international law', including 'respect for intellectual property' and 'fundamental human rights to freedom of expression and association' (Cabinet Office 2011, 27). Australia's 2016 *Cyber Security Strategy: Enabling Innovation, Growth & Prosperity* advocated 'an open, free and secure Internet based on our values of freedom of speech, right to privacy and rule of law' and a preference for multi-stakeholder governance of the

---

5   This excludes 'departmental' cyber security strategies published by the US Department of Defence, Department of Homeland Security, etc.

internet, with the fundamental belief that 'state behaviour in cyberspace is governed by international law' (Commonwealth of Australia 2016, 41). The UK's *National Cyber Security Strategy 2016–2021* also promoted 'the application of international law in cyberspace' as well as 'the agreement of voluntary, non-binding norms of responsible state behaviour' (Cabinet Office 2016, 49). In 2017, Australia published its own *International Cyber Engagement Strategy*, outlining Australia's understanding of international law's applicability to 'state conduct in cyberspace' (mostly based on the 2012–2013 UNGGE report) and the country's position on norms of responsible state behaviour in cyberspace (mostly based on the 2014–2015 UNGGE report) (DFAT 2017, 90–94).

In May 2018, the UK outlined its understanding of international law's applicability to cyberspace through a speech delivered by its attorney general (Wright 2018). It specifically highlighted the importance of the UN Charter and Article 2(7) (prohibiting interventions in the domestic affairs of states), Article 2(4) (prohibiting 'the threat or use of force against the territorial independence or political integrity of any state'), and Article 51 (the right to self-defence if cyber operations result in or present an imminent threat of 'death and destruction on an equivalent scale to an armed attack'). On foreign interference, the speech argued that if hostile states used cyber operations 'to manipulate the electoral system to alter the results of an election' or intervened 'in the fundamental operation of Parliament' or the stability of financial systems, that would 'surely be a breach of the prohibition on intervention in the domestic affairs of states'. In the context of the cyberspace 'Great Game', it is hardly surprising that the examples of foreign interference used by the attorney general described the kinds of activities Russia and China have been regularly accused of undertaking in Five Eyes countries (DHS 2016; Packham 2019). Finally, and rather controversially given Sino-Russian concerns over the militarization of cyberspace, the speech argued that 'each state has the right to develop a sovereign offensive cyber capability', which would not, however, imply the militarization of cyberspace because states had 'an obligation' to ensure such capabilities were used 'in accordance with international law' (Wright 2018).

As well as shaping the governance of cyberspace through inputs into the UNGGE, the Five Eyes also promote several highly prominent norm building initiatives they helped establish: the Council of Europe's *Convention on Cybercrime* (the Budapest Convention), the North Atlantic Treaty Organization (NATO)'s Cooperative Cyber Defence Centre of Excellence (CCDCOE), the Global Conference on Cyberspace and the Global Commission on the Stability of Cyberspace (GCSC), and the Freedom Online Coalition (FOC).

The Budapest Convention is arguably the most prominent international treaty outlining specific practices for combating transnational cyber crime (Council of Europe n.d.). All Five Eyes countries aside from New Zealand[6] are parties to it and regularly promote its virtues in fighting cyber crime. While the Convention's stipulation on trans-border access to data without the need to consult national governments (Article 32) provides for a transnationalism in line with Five Eyes cyberspace norm preferences, it clashes with the Sino-Russian preference for cyber sovereignty (China and Russia are not parties to the Convention).

NATO CCDCOE is best known for publishing the *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*, which is promoted as the 'most comprehensive analysis on how existing international law applies to cyberspace' (CCDCOE n.d.). While the manual does 'not reflect official NATO opinion', it is sponsored by NATO, the US Cyber Command, and the International Committee of the Red Cross and generally underpins the Five Eyes' promotion of the applicability of existing international law in cyberspace (Whyte and Mazanec 2019, 255).

The Global Conference on Cyberspace, which grew out of the 2011 London Conference on Cyberspace, currently represents one of the most prominent international cyber security forums (Cabinet Office 2011, 27). It has produced the 2013 *Seoul Framework for and Commitment to an Open and Secure Cyberspace*, a set of cyberspace norm preferences favoured by the Five Eyes and like-minded states. The Framework promotes international law's applicability to cyberspace, respect for human rights online, multi-stakeholder management of the internet, and for states to 'meet their international obligations regarding internationally wrongful acts attributable to them' (Seoul Framework 2013). The 2015 Global Conference on Cyberspace led to the formation of the Global Commission on the Stability of Cyberspace (GCSC), which in 2019 published its own set of cyberspace norms, promoting the 'integrity of the public core of the Internet' and proscribing foreign interference in 'elections, referenda or plebiscites' (GCSC 2019, 8–9). While the Global Conference on Cyberspace and the GCSC are made up of international delegates and commissioners, the Sino-Russian bloc perceives these two initiatives as dominated by the Five Eyes and like-minded Western states (Segal 2017, 8–9). Therefore, in 2014 China inaugurated its own international forum for promoting Sino-Russian cyberspace norm preferences, the World Internet Conference (World Internet Conference n.d.).

Finally, the Freedom Online Coalition (FOC), whose membership is almost exclusively Western and European, is 'committed to advancing internet freedom – free expression, association, assembly, and privacy online – worldwide' (FOC n.d.). It holds regular conferences, as well as meetings at the margins of UN cyber-dedicated

---

6   The New Zealand government has received legislative advice to become a party to the Budapest Convention or to amend existing national legislation to be in line with it. See New Zealand Law Commission and Ministry of Justice 2015, 28, 190, 207–210, 261–264.

fora, and is regularly promoted by Five Eyes countries (New Zealand Government 2015, 10; Commonwealth of Australia 2016, 40–41; US 2018, 11, 25). The FOC has also launched a Digital Defenders Partnership (DDP) to 'protect critical Internet users' like 'human rights defenders' (activists, bloggers, civil society organisations, and journalists) 'to defend human rights, and keep the Internet free and open' (DDP n.d.).

B. The Sino-Russian cyber security norm preferences

Notwithstanding their nominal agreement with the Five Eyes about existing international law's applicability to cyberspace, China and Russia are not part of the previously mentioned cyberspace norm building initiatives. The Sino-Russian bloc perceives the Budapest Convention, the Tallinn Manual, the GCSC, and the FOC as part of a Western-centric norm building system, infused with the global liberal agenda underpinning and informing the foreign and strategic policies of the Five Eyes. While the Sino-Russian bloc favours new international treaties, rather than the development of non-binding norms, to govern cyberspace, it still participates in developing cyberspace norms mainly through existing multilateral platforms like the UN and the Shanghai Cooperation Organisation (SCO).[7]

Sino-Russian cyberspace norm preferences have three key features. First and foremost is the conceptual difference between 'information security' and 'cyber security'. For the UK (Cabinet Office 2016, 15), cyber security entails 'the protection of information systems (hardware, software and associated infrastructure), the data on them, and the services they provide, from unauthorised access, harm or misuse', while for Australia (DFAT 2017, 23), cyber security enables 'access to online information by individuals, governments and businesses, while ensuring the information and the systems that underpin it are protected from unauthorised access, removal or change'. Hence, for the Five Eyes, cyber security is primarily about the integrity of the systems delivering the information, and only by extension the information itself. By contrast, the Sino-Russian definition of information security entails 'the status of individuals, society and the state and their interests when they are protected from threats, destructive and other negative impacts in the information space' (SCO 2009) and the 'practice of defending the information of individuals, society and the government from unauthorized access, use, disclosure, disruption, modification, perusal, inspection, recording or destruction' (CSIS 2015). Hence, for the Sino-Russian bloc, information security is primarily about the information itself, although the integrity of the systems delivering that information is also crucial. This concept allows governments more arbitrary leeway to interpret what constitutes 'threats, destructive and other negative impacts in the information space'.

---

[7] The SCO is an intergovernmental organization founded in 2001 by China, Russia, Kazakhstan, Kyrgyzstan, Tajikistan, and Uzbekistan (and joined in 2017 by India and Pakistan).

The preference for information security underpins the second central feature of Sino-Russian cyberspace norm preferences, that of cyber sovereignty – a feature consistent with broader Sino-Russian foreign and strategic policy concerns regarding the Five Eyes' global liberal agenda. The SCO's 2009 *Agreement on Cooperation in Ensuring International Information Security between the Member States of the Shanghai Cooperation Organization* identified 'information weapons', 'information warfare', and the 'dissemination of information prejudicial to the socio political and socio economic systems, spiritual, moral and cultural environment of other States' as key threats, highlighting 'non-interference' as the key principle of international information security cooperation (Articles 2, 4). In 2016, China's first *National Cyberspace Security Strategy* argued that cyberspace was a new domain 'for national sovereignty', identifying the protection of 'sovereignty in cyberspace' as a key principle (CCM 2016). The Strategy identified 'cyber penetration' and the 'use of networks to interfere in the internal political affairs of other countries' as the foremost 'grave challenge' facing China in cyberspace. It also recognized the growing international 'competition in cyberspace', with its 'strife for the control of strategic resources', as a key cyberspace challenge, taking a swipe at the Five Eyes' dedication to 'cyber deterrence' by warning that 'a small number of countries is strengthening a cyber deterrence strategy, aggravating an arms race in cyberspace, and bringing new challenges to global peace'. 'Resolutely defending sovereignty in cyberspace' was and still is China's primary strategic task.

In April 2015, China and Russia's agreement 'on cooperation in ensuring international information security' formalized the Sino-Russian cyber security bloc. The agreement made clear the bloc's primary concern with cyber sovereignty by emphasizing that 'the state has the sovereign right to define and implement public policies on matters relating to' ICT and the internet (CSIS 2015, 4). Article 2 of the agreement lists the six 'main threats in the field of international information security'. The first, second, fifth, and sixth are all concerned with some form of violating cyber sovereignty; the other two are cyber terrorism and cyber crime. While the agreement's main focus may be on cooperation on internet control to help maintain domestic stability (Segal 2020), Articles 3.3 and 3.13 explicitly discuss 'cooperation in the development and promotion of international law in order to ensure national and international information security' and enhancing 'cooperation and coordination' on 'issues of international information security within the framework of international organizations and fora'.

China and Russia's desire to challenge what they perceive to be the Five Eyes' dominance in shaping how cyberspace is governed is the third key feature of Sino-Russian cyberspace norm preferences. Much like the focus on cyber sovereignty, this feature is underpinned by China and Russia's broader foreign and strategic policies, which seek to reform the post-World War II Five Eyes-dominated global security

order to their advantage. China is keen on reforming the Internet's governance, arguing that since cyberspace 'is the common space of activities for mankind', it should be governed following 'a multilateral approach' whereby 'countries, big or small, strong or weak, rich or poor, are all equal members of the international community entitled to equal participation in developing international order and rules in cyberspace' (MFAPRC 2017). China states that the UN 'should play a leading role in coordinating positions of various parties and building international consensus' on the internet's governance, arguing for a 'multilateral' model of internet governance that gives greater control to governments and political regimes. The Sino-Russian internet governance reform agenda is wide, seeking to enhance the UN's role by reforming the Internet Corporation for Assigned Names and Numbers (ICANN) and replacing the Budapest Convention with a new UN cyber crime treaty. China will 'push for institutional reform of the UN Internet Governance Forum to enable it to play a greater role in Internet governance' and 'vigorously promote the reform of ICANN to make it a truly independent international institution, increase its representations and ensure greater openness and transparency in its decision-making and operation' (MFAPRC 2017). Russia views the Budapest Convention's transnationalism as violating 'principles of state sovereignty and non-interference' and has won support at the UN for a 'committee of experts' to consider the development of a new cyber crime treaty to replace it (Segal 2020).

The Sino-Russian bloc has primarily used the UN and SCO to promote its vision of the governance of cyberspace and cyberspace norm preferences. The aforementioned 2009 SCO Agreement made clear the Sino-Russian primary concern with information security and full control of data within a state's territory. Building on this, in September 2011, China, Russia, Tajikistan, and Uzbekistan submitted to the UN General Assembly their proposal for a voluntary *International Code of Conduct for Information Security*. It outlined cyberspace norms which, among other issues, called on states to comply with the UN Charter and respect the 'sovereignty, territorial integrity and political independence' of other states; abstain from using ICT for hostile activities and 'acts of aggression'; prevent states from using their ICT advantages 'to threaten the political, economic and social security of other countries'; promote 'multilateral' governance of the internet; and settle disputes peacefully, refraining from 'the threat or use of force'. In highlighting the primacy of cyber sovereignty, the Code affirmed the rights of states to 'protect, in accordance with relevant laws and regulations, their information space and critical information infrastructure from threats, disturbance, attack and sabotage' (UN A/66/359 [2011], 4–5). Finally, in January 2015, the Sino-Russian bloc submitted to the UN General Assembly a revised *International Code of Conduct for Information Security*, which reiterated their focus on cyber sovereignty, reaffirming the rights of states to protect their 'information space and critical information infrastructure against damage resulting from threats, interference, attack and sabotage' (UN A/69/723 [2015], 4–6).

4. CONCLUSION

The development of norms of responsible state behaviour in cyberspace is fundamentally a political process, underpinned by values, ideologies, and interests inherent in a state's foreign and strategic policy considerations. States do not seek to neutrally shape norms in cyberspace for the sake of some abstract universal good but rather to expand their own ideological preferences and values and advance their own 'economic and security interests' (Cabinet Office 2016, 9; US 2018, 3). Therefore, cyberspace norm building entails the contestation and competition of ideas, values, and interests inherent in 'regular' international relations – it is the continuation of foreign and strategic policy by other means. All of this is visible in the cyberspace 'Great Game', in which the Five Eyes and the Sino-Russian bloc compete to dominate the governance of cyberspace and global cyberspace norms. Two fundamentally different visions, underpinned by irreconcilable political ideologies, values, and interests promoted by the world's greatest cyber powers, highlight the overwhelming importance and seriousness of this competition.

Both the Five Eyes and the Sino-Russian bloc agree that existing international law applies to cyberspace, but their approaches to the governance of cyberspace exhibit substantive conceptual differences. The Five Eyes' preference for 'cyber security' and the Sino-Russian preference for 'information security' place primary focus on two different issues: securing the infrastructure and processes through which information in cyberspace is accessed versus securing the very information that is being accessed. The Five Eyes' preference for a 'multi-stakeholder' governance of the internet and the Sino-Russian preference for a 'multilateral' approach are also different: the former emphasizes the role of non-governmental actors, staying true to the internet's diffuse origins and operational nature, while the latter emphasizes the primacy of governments and state officials in a more centralized approach. Finally, although the Five Eyes' preference for 'transnationalism' (open internet) and the Sino-Russian preference for 'cyber sovereignty' is not underpinned by a substantive conceptual difference, the significant difference in emphasis placed on this issue has made it arguably the most fundamental point of contention between the two blocs. While both blocs subscribe to the notion of cyber sovereignty, the Five Eyes emphasize the virtues of an open internet and a transnational approach to data management (partially because of their dominance in the development of cyberspace), while the Sino-Russian bloc emphasizes the virtues of territorial integrity and sovereignty (partially because of their weariness of, and desire to challenge, the Five Eyes' dominance in cyberspace).

The implications of the 'Great Game' for the governance of cyberspace are significant. The contested and competitive nature of the 'Great Game' has entrenched 'norm siloing', whereby like-minded states with shared ideologies, values, and interests

band together to establish and promote favoured cyberspace norm preferences, 'leading to competing or conflicting islands of normality' (Finnemore and Hollis 2016, 466). However, while such cyberspace norm building may be easier to achieve with like-minded states, those states are usually not the ones whose malicious cyber activity such norms aim to constrain. Moreover, the 'rise of digital authoritarianism' and the decline of global internet freedoms suggests that the Sino-Russian bloc may be slowly gaining the upper hand in the cyberspace 'Great Game' (Shahbaz 2018). The widespread concern for the survival of political regimes with varying degrees of authoritarianism and patrimonialism, coupled with low governmental levels of cyber capability and resources, will make many UN member states fundamentally more sympathetic to the Sino-Russian preference for information security and cyber sovereignty. Although the Five Eyes are powerful enough to maintain their cyber dominance for the foreseeable future, to counter the Sino-Russian bloc they will probably reinvigorate efforts to promote their vision of cyber security and a free and open internet. This will increase the competitiveness and pitfalls of failure in the cyberspace 'Great Game', with neither side likely to 'win', because the appeal of their cyberspace norm preferences will vary depending on the extent to which other states perceive those norms as compatible (or incompatible) with their own ideologies and values, and helpful (or unhelpful) in fulfilling their own wider foreign policy and strategic interests. The cyberspace 'Great Game' will remain a defining feature of global efforts to govern cyberspace in the 2020s, and it is highly unlikely that the latest UNGGE session (2019–2021) will change that in any substantive way.

REFERENCES

Barrinha, Andrew, and Thomas Renard. 2020. 'Power and Diplomacy in the Post-Liberal Cyberspace'. *International Affairs* 96, no. 3: 749–766.

Bolt, Paul J., and Sharyl N. Cross. 2018. *China, Russia, and Twenty-First Century Global Geopolitics*. Oxford: Oxford University Press.

Cabinet Office. 2011. *The UK Cyber Security Strategy: Protecting and Promoting the UK in a Digital World*. https://www.gov.uk/government/publications/cyber-security-strategy.

Cabinet Office. 2016. *National Cyber Security Strategy 2016–2021*. https://www.gov.uk/government/publications/national-cyber-security-strategy-2016-to-2021.

Center for Strategic and International Studies (CSIS). 2015. *Sino-Russian Cybersecurity Agreement 2015*. https://www.csis.org/blogs/strategic-technologies-blog/sino-russian-cybersecurity-agreement-2015.

China Copyright and Media (CCM). 2016. *National Cyberspace Security Strategy*. https://chinacopyrightandmedia.wordpress.com/2016/12/27/national-cyberspace-security-strategy/.

Clausewitz, Carl von. 2007. *On War*. Oxford: Oxford University Press.

Commonwealth of Australia. 2016. *Australia's Cyber Security Strategy: Enabling Innovation, Growth & Prosperity*. https://www.homeaffairs.gov.au/cyber-security-subsite/files/PMC-Cyber-Strategy.pdf.

Cooperative Cyber Defence Centre of Excellence (CCDCOE). n.d. *About Us/History*.
https://ccdcoe.org/about-us/.

Council of Europe. n.d. *Convention on Cybercrime*.
https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/185.

Cox, James. 2012. 'Canada and the Five Eyes Intelligence Community'. *OpenCanada.Org*. 18 December 2012.
https://opencanada.org/canada-and-the-five-eyes-intelligence-community/.

Department of Foreign Affairs and Trade (DFAT). 2017. *Australia's International Cyber Engagement Strategy*.
https://www.dfat.gov.au/publications/international-relations/international-cyber-engagement-strategy/aices/
index.html.

Department of Homeland Security (DHS). 2016. *Joint Statement from the Department Of Homeland
Security and Office of the Director of National Intelligence on Election Security*. https://www.dhs.gov/
news/2016/10/07/joint-statement-department-homeland-security-and-office-director-national.

Digital Defenders Partnership (DDP). n.d. *About*. https://www.digitaldefenders.org/about/.

Duckett, Chris. 2020. 'Canadian Major Telcos Effectively Lock Huawei out of 5G Build'. ZDNet. 3 June 2020.
https://www.zdnet.com/article/canadian-major-telcos-effectively-lock-huawei-out-of-5g-build/.

Finnemore, Martha, and Duncan B. Hollis. 2016. 'Constructing Norms for Global Cybersecurity'. *American
Journal of International Law* 110, no. 3: 425–479.

Freedom Online Coalition (FOC). n.d. *Freedom Online Coalition: Factsheet*.
https://freedomonlinecoalition.com/about-us/members/.

Global Commission on the Stability of Cyberspace (GCSC). 2019. *Advancing Cyberstability*. The Hague Centre
for Strategic Studies and EastWest Institute. https://cyberstability.org/report/.

Global Change Data Lab (GCDL). n.d. 'Share of the Population Using the Internet, 1990 to 2017'.
https://ourworldindata.org/grapher/share-of-individuals-using-the-internet?tab=chart&country=AUS~CAN
~USA~GBR~NZL.

Gold, Hadas. 2020. 'UK Bans Huawei from its 5G Network in Rapid About-Face'. *CNN*. 14 July 2020.
https://edition.cnn.com/2020/07/14/tech/huawei-uk-ban/index.html.

Legrand, Tim. 2019. 'The Past, Present and Future of Anglosphere Security Networks: Constitutive Reduction of
a Shared Identity'. In *The Anglosphere: Continuity, Dissonance and Location*, edited by Ben Wellings and
Andrew Mycock, 56–76. Oxford: Oxford University Press.

Lilly, Bilyana, and Joe Cheravitch. 2020. 'The Past, Present, and Future of Russia's Cyber Strategy and
Forces'. In *12th International Conference on Cyber Conflict 20/20 Vision: The Next Decade*, edited by T.
Jančárková, L. Lindström, M. Signoretti, I. Tolga, and G. Visky, 129–155. NATO: CCDCOE Publications.

Lukin, Alexander. 2018. *China and Russia: The New Rapprochement*. Cambridge: Polity Press.

Ministry of Foreign Affairs of the People's Republic of China (MFAPRC). 2017. *International Strategy of
Cooperation on Cyberspace*. https://www.fmprc.gov.cn/mfa_eng/wjb_663304/zzjg_663340/jks_665232/
kjlc_665236/qtwt_665250/t1442390.shtml.

Moon, Louise, and Chad Bray. 2019. 'Donald Trump's Huawei Ban is a More Severe Threat to Global Economy
than Trade War Tariffs, Economists Say'. *South China Morning Post*. 24 May 2019. https://www.scmp.
com/business/companies/article/3011676/trumps-huawei-ban-more-severe-threat-global-economy-trade-
war.

National Security Agency Central Security Service (NSACSS). n.d. *UKUSA Agreement 1956*. https://www.nsa.
gov/News-Features/Declassified-Documents/UKUSA/.

New Zealand Government. 2015. *New Zealand's Cyber Security Strategy 2015 Action Plan*. https://www.itu.int/en/ITU-D/Cybersecurity/Documents/National_Strategies_Repository/nz-cyber-security-action-plan-december-2015.pdf.

New Zealand Law Commission and Ministry of Justice. 2015. *Review of the Search and Surveillance Act 2012*. Report 141. https://www.lawcom.govt.nz/sites/default/files/projectAvailableFormats/NZLC-R141-Review-of-the-Search-and-Surveillance-Act-2012-final_0.pdf.

Packham, Colin. 2019. 'Exclusive: Australia Concluded China Was Behind Hack on Parliament, Political Parties – Sources'. *Reuters*. 16 September 2019. https://www.reuters.com/article/us-australia-china-cyber-exclusive-idUSKBN1W00VF.

Richelson, Jeffrey T. 2016. *The U.S. Intelligence Community*. 7th edition. Boulder: Westview Press.

Segal, Adam. 2017. *Chinese Cyber Diplomacy in a New Era of Uncertainty*. Hoover Working Group on National Security, Technology, and Law: Aegis Paper Series No. 1703. https://www.hoover.org/research/chinese-cyber-diplomacy-new-era-uncertainty.

Segal, Adam. 2020. 'Peering into the Future of Sino-Russian Cyber Security Cooperation'. *War on the Rocks*. 10 August 2020. https://warontherocks.com/2020/08/peering-into-the-future-of-sino-russian-cyber-security-cooperation/.

Seoul Framework for and Commitment to Open and Secure Cyberspace (Seoul Framework). 2013. https://www.un.org/disarmament/wp-content/uploads/2019/10/ENCLOSED-Seoul-Framework-for-and-Commitment-to-an-Open-and-Secure-Cyberspace.pdf.

Shahbaz, Adrian. 2018. 'The Rise of Digital Authoritarianism. Freedom on the Net 2018'. *Freedom House*. https://freedomhouse.org/report/freedom-net/2018/rise-digital-authoritarianism.

Shanghai Cooperation Organization (SCO). 2009. *Agreement on Cooperation in Ensuring International Information Security between the Member States of the Shanghai Cooperation Organization*. http://eng.sectsco.org/documents/.

Slezak, Michael, and Ariel Bogle. 2018. 'Huawei Banned from 5G Mobile Infrastructure Rollout in Australia'. *ABC News*. 23 August 2018. https://www.abc.net.au/news/2018-08-23/huawei-banned-from-providing-5g-mobile-technology-australia/10155438.

Tikk, Eneken, and Mika Kerttunen. 2017. *The Alleged Demise of the UN GGE: An Autopsy and Eulogy*. New York: Cyber Policy Institute. https://cpi.ee/wp-content/uploads/2017/12/2017-Tikk-Kerttunen-Demise-of-the-UN-GGE-2017-12-17-ET.pdf.

Tobin, Meaghan. 2019. 'New Zealand Bans Huawei from 5G, China Has Message for New Zealand'. *South China Morning Post*. 17 February 2019. https://www.scmp.com/week-asia/geopolitics/article/2186402/new-zealand-bans-huawei-china-has-message-new-zealand.

Trump, Donald J. 2019. *Executive Order on Securing the Information and Communications Technology and Services Supply Chain*. 15 May 2019. https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-securing-information-communications-technology-services-supply-chain/.

UN A/Res/65/201 (2010). *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*. United Nations General Assembly.

UN A/66/359 (2011). *Annex to the Letter Dated 12 September 2011 from the Permanent Representatives of China, the Russian Federation, Tajikistan and Uzbekistan to the United Nations Addressed to the Secretary-General*. United Nations General Assembly.

UN A/Res/68/98 (2013). *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*. United Nations General Assembly.

UN A/Res/70/174 (2015). *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*. United Nations General Assembly.

UN A/69/723 (2015). *Annex to the Letter Dated 9 January 2015 from the Permanent Representatives of China, Kazakhstan, Kyrgyzstan, the Russian Federation, Tajikistan and Uzbekistan to the United Nations Addressed to the Secretary General*. United Nations General Assembly.

United Nations Institute for Disarmament Research (UNIDIR). 2016. *Report of the International Security Cyber Issues Workshop Series*. https://unidir.org/publication/report-international-security-cyber-issues-workshop-series.

United States (US). 2011. *International Strategy for Cyberspace*. https://obamawhitehouse.archives.gov/sites/default/files/rss_viewer/international_strategy_for_cyberspace.pdf.

United States (US). 2018. *National Cyber Strategy of the United States of America*. https://trumpwhitehouse.archives.gov/wp-content/uploads/2018/09/National-Cyber-Strategy.pdf.

Voo, Julia, Irfan Hemani, Simon Jones, Winnona DeSombre, Daniel Cassidy, and Anina Schwarzenbach. 2020. *National Cyber Power Index 2020*. China Cyber Policy Initiative: Belfer Center for Science and International Affairs.

Vucetic, Srdjan. 2010. 'Anglobal Governance?' *Cambridge Review of International Affairs* 23, no. 3: 455–474. https://doi.org/10.1080/09557570903535755.

Wellings, Ben, and Andrew Mycock, eds. 2019. *The Anglosphere: Continuity, Dissonance and Location*. Oxford: Oxford University Press.

Whyte, Christopher, and Brian Mazanec. 2019. *Understanding Cyber Warfare: Politics, Policy and Strategy*. New York: Routledge.

World Internet Conference. n.d. http://www.wuzhenwic.org/.

Wright, Jeremy. 2018. *Cyber and International Law in the 21st Century*. https://www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century.

The Global Spread of Cyber Forces, 2000–2018

Jason Blessing

DAAD Post-Doctoral Fellow[1]

Foreign Policy Institute

Johns Hopkins School of Advanced International Studies

Washington, DC, United States

jblessing@jhu.edu

Abstract: Although militaries have been building cyber capabilities since the late 1980s, formalized military organizations for these capabilities have only recently emerged. These cyber forces—active-duty military organizations that possess the capability and authority to direct and control cyberspace operations for strategic ends—have spread rapidly across the international system since the first few years of the 21st century. This article catalogues the development of cyber forces across the globe and assesses the various force structures. Existing research has largely been confined to examinations of cyber forces in North Atlantic Treaty Organization (NATO) member states. This article provides a broader view of global developments by introducing new data on the worldwide spread of cyber forces from 2000 to 2018. It also offers a typology for assessing cyber force structure based on both organizational model (branch, service, or joint model) and the scale of command (subordinated, sub-unified, or unified). As a result, this article identifies nine distinct cyber force structures. Empirical analysis reveals that 61 United Nations member states had created a cyber force by 2018. Contrary to conventional expectations, this analysis shows increasing variation in cyber force structure over time; no dominant organizational model or force structure has emerged.[2]

Keywords: *cyber forces, cyber force structure, military organizations, cyberspace operations*

---

1. INTRODUCTION

Militaries have been building cyber capabilities since the late 1980s (Wiener 2016); however, formalized military organizations for these capabilities have only recently emerged. The United States Cyber Command, created in 2010 and elevated to an independent unified combatant command in 2017, stands as an obvious example. A variety of other states have also established their own "cyber commands," including South Korea in 2010; Colombia in 2011; the United Kingdom, Turkey, and Spain in 2013; and the Netherlands and Ecuador in 2014 (Keck 2014; *Dialogo* 2013; Osula 2015; Seker and Tolga 2018; Cendoya 2016; Kaska 2015; Directorate of Social Communication of the Joint Command of the Armed Forces of Ecuador 2015).

To date, systematic research on cyber forces[3] has focused more on evaluating organizational maturity than on assessing variations in force structure (for example, see Robinson et al. 2013; Smeets 2019). Extant studies of force structure have tended to examine individual cases like the United States, Russia, China, and North Korea (Nielsen 2016; Lilly and Cheravitch 2020; Costello and McReynolds 2018; Kong et al. 2019). Research in a comparative context has been rare (Gorwa and Smeets 2019). As a notable exception, Pernik (2018) identifies three types of cyber forces: divisions under logistical branches; standalone combat services; and independent combatant commands/branches. Although Pernik (2018) is the first study to explicitly compare organizational arrangements, it captures only a fraction of the possible variation in force structure. Its scope is also limited to five European states, with Finland the only non-NATO state examined.

The lack of extensive comparative research on cyber force structure is problematic for at least two reasons. First, many expectations regarding military organizations are rooted in assumptions about the competitive or normative emulation of dominant paradigms (Resende-Santos 2007; DiMaggio and Powell 1983). For example, as militaries grappled with air power, an independent air force gradually emerged as the dominant organizational paradigm over other alternatives like separate air wings for each service (Hasik 2016). Such expectations leave scholars and practitioners without an appropriate terminology for understanding cyber forces. Indeed, referring to all institutional arrangements as "cyber commands" masks important variation in the scope, roles, and responsibilities in cyberspace. Second, force structures and the organizational origins of cyber forces can shape behavior and the tradeoff between exploitation and disruption. For instance, cyber forces originating in combat services may be more predisposed to take overt military action in cyberspace than those emerging from military intelligence, which may prefer information collection and covert operations (Schneider 2019, 115–120).

---

[3]   Some works use the terms "military cyber organization" or "cyber command." This paper uses "cyber force," since "cyber force structure" is more concise than "military cyber organization force structure" and more precise than "cyber command structure."

This article builds on the works cited above and previously unpublished work (Blessing 2020b) to offer a broader perspective by cataloging the global development of cyber forces. Accordingly, the paper introduces a new database on cyber force structures and examines trends across states both in and outside the North Atlantic Treaty Organization. This paper also advances a novel typology of force structure that provides a foundation for addressing questions related to organizational structure and strategic behavior in cyberspace.

This paper proceeds in five sections. Section 2 defines cyber forces, while Section 3 provides a novel framework for distinguishing cyber forces structures based on organizational model (branch, service, joint) and scale of command (subordinated, sub-unified, unified). Section 4 presents a new database on the global spread of cyber forces from 2000 to 2018. Analysis reveals that 61 United Nations member states had created a cyber force by 2018. The data also show increasing variation in force structure over time; no dominant cyber force structure has emerged. Section 5 explores the implications for NATO, while Section 6 concludes by summarizing and considering future research.

2. DEFINING CYBER FORCES

Existing works describe cyber forces as a kind of military organization with some degree of authority over cyber operations. Pernik (2018, 2–3) states that the term "cyber force" "generally denotes a standalone structure, branch, or service of the armed forces that directs and controls the three main categories of cyberspace operations [defense, exploitation, attack]." Similarly, Smeets (2019, 165) defines a cyber force as "a command, service, branch, or unit within a government's armed forces which has the authority and mission to conduct offensive cyber operations to disrupt, deny, degrade and/or destroy."

Yet not all cyber forces will have the mandate over the full spectrum of operations (as advanced by Pernik) or the full capability to undertake offensive operations (as laid out by Smeets). Moreover, these definitions are generally agnostic as to the strategic ends pursued by cyber forces. A key problem for distinguishing force structures, then, is determining which organizations are excluded.

This article defines cyber forces as active-duty military organizations with the capability and authority to direct and control strategic cyberspace operations to influence strategic diplomatic and/or military interactions (on cyberspace operations and strategic interactions, see Valeriano and Maness 2015). Cyberspace operations can include defense to prevent the compromise of the integrity, confidentiality, or

availability of information on computers, the computers themselves, or networks; exploitation to collect information from an adversary's computers and networks that fall short of disrupting or destroying information; and attacks to disrupt, deny, degrade, or destroy information on computers or the computers or networks themselves. Espionage and theft constitute attacks when information or systems are destroyed (Healey 2013, 279–280).

This definition excludes three types of organizations with similar missions. The first is civilian intelligence agencies like the U.S. National Security Agency. Despite potentially significant overlaps in operations, the primary purposes of civilian agencies and cyber forces are fundamentally different. Aside from falling outside military chains of command, civilian intelligence agencies are largely focused on information collection. While cyber forces can and do collect information, intelligence-gathering is generally in service of and subordinated to gaining strategic advantage.

Second, purely reservist components—like Estonia's Cyber Defense Unit and Latvia's Cyber Defense Unit (Gramaglia et al. 2013; Gelzis 2014)—are excluded. Although reservists can provide several benefits (Miller et al. 2013; Baezner 2020), reservist units cannot maintain full-time authority over cyberspace operations. Reservist operation is conditional on legal activation (Brenner and Clarke 2011), and many governments maintain restrictions on using reserve funds for operational missions. Reservist units are also highly fluid: they consist of volunteers serving for only limited periods. This fluidity can compromise the up-to-date knowledge of operations, scalability, and interoperability required of active-duty organizations (Applegate 2012; Curley 2018). Overcoming such challenges would require substantial volunteering past minimum requirements, an assumption unlikely to hold across militaries.

Finally, military computer emergency readiness teams (MilCERTs), incident response teams (MilCIRTs), and incident response centers (MilCIRCs) are excluded. These organizations—like the Jordanian Armed Forces' MilCERT and Moldovan Armed Forces' MilCIRC (North Atlantic Treaty Organization 2017; de Albuquerque and Hedenskog 2016)—look for and patch military and/or defense network vulnerabilities, develop plans to deal with network outages and malicious attacks, and coordinate responses (Healey 2013, 279). They work defensively at the tactical level to ensure network operability but do not seek to integrate capabilities on larger operational or strategic scales. While they can be under the control of/report to cyber forces, they do not constitute cyber forces.

3. A FRAMEWORK FOR CYBER FORCE STRUCTURE

Like traditional force structure, cyber force structure is crucial for understanding how militaries translate material and human strengths into power on the battlefield. Force structure conventionally refers to the number and types of combat units a military can generate and sustain. It can be defined in several ways: the composition and structure of organizations; unit functions; capabilities; costs of operation; or some combination of these factors (Congressional Budget Office 2016). Unfortunately, much of these data for cyber forces—like personnel costs, operating costs, and capability acquisitions—are either inconsistently documented or remain classified.

This paper proposes two criteria for categorizing cyber force structures: organizational model and scale of command. These dimensions provide important insight into cyber forces' internal organization and how they relate to command structures across the military's combat and combat support subsystems (on militaries as organizations with subsystems, see Farrell 1996). Organizational model helps define combat service membership, internal divisions of labor, and how the cyber force relates to other military components (Augier et al. 2015). Scale of command illuminates the delegation of authority and responsibilities in military hierarchies. Thus it helps assess how operations are coordinated and/or integrated across other mission areas (Brooks 2006, 405–407).

There are three potential organizational models: a branch, service, or joint model. Under a *branch* model, authority for cyberspace operations rests primarily in logistical branches, military intelligence agencies, or signals corps within the combat support subsystem. While combat services can provide personnel for staffing, branch model forces fall outside service department chains of command.[4] Accordingly, a branch model arranges personnel along functional lines—specific expertise, tools, or missions—and not service-based ones. Cyber forces are organized according to a *service* model when a single combat service—domain-based (army, navy, air force) or functional (such as rocket forces, marines, or other standalone services)—retains primary authority for cyberspace operations. In these instances, a cyber force is staffed only by personnel from the combat service to which it reports. Like a branch model, service model personnel are generally grouped according to functional expertise in units or component commands. A *joint* model entails the shared distribution of authority across two or more combat services. Under this model, combat services are force providers—cyber forces rely primarily on the services for staffing and funding. Staffing generally occurs on a short-term, rotational basis. In other words, combat services provide personnel for specific periods before they are recalled to service-based assignments and replaced in the cyber force with other service personnel.

---

4   While combat support elements are present within combat subsystems, combat is the overarching functional role for that subsystem.

A joint model thus serves to facilitate coordination among service components. Therefore, service membership is the primary organizing principle within a joint model; functional expertise is a secondary principle.

Cyber forces can also be classified by subordinated, sub-unified, or unified commands. *Subordinated* cyber forces appear when existing commands incorporate cyberspace operations to support ongoing missions and enhance effectiveness without disrupting status quos (on military adaptation, see Farrell 2010). *Sub-unified* force structures consist of specialized sub-organizations that treat cyber operations as an independent mission. These can result from reconfigurations of personnel and capabilities within subsystems to implement novel operational concepts or technologies. *Unified* forces institutionalize "new ways of war" (Rosen 1991) related to the cyber domain via a new branch, service, or combatant command. They can emerge from military-wide reorganizations that disrupt relationships and interdependencies. Unified forces have no parent organization and report directly to chiefs/ministers/secretaries of defense. Sub-unified forces report to existing unified commands; subordinated forces report to sub-unified (and, in rare cases, unified) parent commands.

These two dimensions of force structure have important implications for the functioning and behavior of cyber forces. For example, all else held equal, unified cyber forces with greater scales of command are likely to be better resourced, better staffed, and better positioned to compete for additional resources than sub-unified or subordinated forces. Scale of command also provides a proxy for the development of and degree to which cyber capabilities are considered an independent military tool. The branch, service, and joint models give additional insight into behavior. For instance, because a joint model incorporates multiple service elements, it can facilitate the development of doctrine for multi-domain operations. Yet a joint model must also grapple with service prerogatives and parochialism that can hamper effectiveness. Inter-service competition similarly affects service model cyber forces. And while service models may be able to better develop cyber personnel (through specialized service academy training and new career paths), they risk losing mission independence to existing service priorities. Branch model forces also risk subordination to combat service missions that prevents the development of independent capabilities.

Table I summarizes the nine cyber force structures produced by these criteria. A brief description of the nine force structures with illustrative examples accompanies Table I.

TABLE I: A TYPOLOGY OF CYBER FORCE STRUCTURES

Organizational Model	Scale of Command		
	Subordinated	**Sub-Unified**	**Unified**
Branch	(1) Subordinated Branch	(4) Sub-Unified Branch	(7) Unified Branch
Service	(2) Subordinated Service	(5) Sub-Unified Service	(8) Unified Service
Joint	(3) Subordinated Joint	(6) Sub-Unified Joint	(9) Unified Joint

(1) Subordinated Branch: non-service communications divisions, signals intelligence units, or military intelligence agencies that integrate cyberspace operations into existing command structures. Examples include Israel's Unit 8200, an electronics intelligence unit under the Directorate of Military Intelligence; and Estonia's Strategic Communications Center, a unit under the Support and Signals Battalion until 2018 (Lewis and Neuneck 2013; Osula 2015a).

(2) Subordinated Service: one combat service co-opts the cyber mission into existing electronic warfare, signals, or communications units; no other services have the capability or mandate to conduct cyberspace operations. The Danish Army's 3rd Electronic Warfare Company (2009–2012) and the Philippine Army's Signals Corps (operational in 2016) are examples of service units with primary responsibilities for cyberspace operations (International Institute for Strategic Studies 2013; Felongco 2016).

(3) Subordinated Joint: primarily temporary, issue- or mission-driven task forces or units that coordinate the cyber mission across two or more combat services. A Subordinated Joint force structure does not include major service-level commands as components. Examples include a variety of joint task forces in the United States (2001–2010)[5] and France's Cyber Defense Cell (2011–2015; see Brangetto 2015).

(4) Sub-Unified Branch: new cyber divisions or directorates under military intelligence agencies, communications/information systems agencies, or joint staff support directorates. Examples include the Finnish Cyber Defense Division (2015–present) and the Cyber Security Operations Center under the Belgian Military Intelligence Service (Pernik 2018; Lasoen 2019).

---

5   Joint Task Force – Computer Network Operations (2001–2004), Joint Task Force – Global Network Operations (2004), and Joint Functional Component Command – Network Warfare (2005–2010). U.S. Cyber Command, "U.S. Cyber Command History," n.d., https://www.cybercom.mil/About/History/.

(5) Sub-Unified Service: major commands within a combat service for conducting cyberspace operations that are on par with existing service commands and missions. Although it can be staffed with personnel from other services, this structure is subordinated to only one service. Examples include Nigeria's Cyber Warfare Command (operational in 2018), which consolidates the Army's efforts into a new service command; and Brazil's Cyber Defense Command (2017–present), which incorporates personnel from the Army, Navy, and Air Force but is under the sole authority of the Army (Moury 2017; Omonobi-Abuja 2018).

(6) Sub-Unified Joint: structure that reports to an existing joint unified combatant command, significantly expanding that parent command's scope of operations. Unlike subordinated structures, Sub-Unified Joint structures are necessarily comprised of major commands from at least two services. United States Cyber Command under U.S. Strategic Command (2010–2017) and Italy's Joint Command for Cyberspace Operations under the Joint C4 Defense Command (operational in 2017) fall in this category (Italian Ministry of Defence 2018).

(7) Unified Branch: independent non-combat military branches that hold special armament or equipment to conduct missions in the cyber domain. Examples include Estonia's Cyber Command (2018–present) and Norway's Cyber Defense Force (2012–present) (Estonian Defence Forces, 2018; Ministry of Defense of Norway 2012).

(8) Unified Service structures are cyber-specific combat services (with military departments) that receive the same hierarchical standing as other domain-based services (armies, navies, and air forces). Only China's Strategic Support Force (2016–present) and Germany's Cyber and Information Domain Service (2017–present) utilize this force structure (International Institute for Strategic Studies 2019; Pernik 2018).

(9) Unified Joint: unified combatant commands for cyberspace comprised of at least two service-level component commands. These independent commands report directly to the top defense official. Examples include U.S. Cyber Command (2017–present) and the Netherlands' Defense Cyber Command (2018–present).

4. CYBER FORCES IN THE WORLD, 2000–2018

To assess the global spread of cyber forces, this article uses a custom-created database introduced in Blessing (2020b): the Dataset on Cyber Force Structures (DCFS). This new database catalogues cyber force structures for the 172 United Nations (UN) members with an active military force from 2000 to 2018. An active military force is a necessary precondition: there can be no cyber force without an active military. Accordingly, the DCFS excludes the 21 UN member states that do not maintain active military forces.[6]

The dataset utilizes five types of sources: government publications; reports from think tanks or international organizations; peer-reviewed academic works; articles from international and regional news outlets; and interviews with former policymakers, military officials, industry members, and subject matter experts. Inclusion in the author-coded dataset requires satisfying the following basic criteria:

- A government source identifies an organization responsible for cyberspace operations. Government sources are corroborated by two other resources. Where government sources are unavailable or lack detail, information is derived from three different categories of resources.

- When multiple organizations are responsible for cyberspace operations, cyber forces are coded based on the military hierarchy: organizations higher in the chain of command with operational responsibilities are designated as the primary cyber force. For example, Denmark's cyber force in 2009–2012 was the Army 3rd Electronic Warfare Company; however, because the Offensive Cyber Warfare Unit (established 2012) under the Defense Intelligence Service had fewer links in the chain of command to the joint Defense Command and Minister of Defense, it replaced the Army's unit as the primary cyber force despite the latter's continued operation.

- Organizational model is based on subsystem and the number of combat services providing personnel. Subsystem is coded on the reporting structures of parent commands. For example, Germany's Department of Information and Computer Network Operations, formerly under the Joint Support Service's Strategic Reconnaissance Command, is coded as combat support. Where no parent organization exists (i.e., for unified commands), subsystem rests on whether the force falls under service chains of command or is a non-

---

[6] Because the database was originally presented as part of doctoral dissertation work in Blessing (2020b), the first round of data collection efforts, covering the period between January 2000 and December 2018, concluded in 2019. As such, the data below do not reflect the most up-to-date force structures for each country. A second round of data collection and coding, which will update the DCFS for the 2019–2021 period, is currently underway and is scheduled to be completed in early 2022.

service force. Cyber forces in combat support subsystems are branch model. Cyber forces in combat subsystems are either service model (one service) or joint model (two or more services). Joint models occur when services are formally linked by a supra-command or maintain independent cyber forces. When multiple services have cyber forces that report to only one service, a service model is coded.

- Scale of command is determined by immediate parent organizations and reporting structures. Unified commands have no parent organizations and report to chiefs/ministers/secretaries of defense. Unified commands are joint combatant commands, independent combat services, or independent branch commands. Sub-unified commands report to unified commands; they encompass joint component commands, combat service major commands, and major commands reporting to an independent branch. Subordinated commands report to sub-unified commands (and, in rare cases, unified commands); they appear as task forces, joint units under component or combatant commands, units in a service-level major command, or functional branch units.

Each observation in the dataset thus contains the following descriptive information: country name; the name of the organization with authority over cyberspace operations as it appears in the military hierarchy; an operational start date (month/year) indicating initial operating capability; an operational end date (month/year) indicating when the organization was disbanded based on expansion, reorganization, and consolidation, or replacement with new initiatives that change the military hierarchy; the parent command to which the organization directly reports; the organization's location in either the combat or combat support subsystem; and the number of combat services staffing the organization.

Figures 1 and 2 chart the development of cyber forces within NATO countries and in the rest of the world between 2000 and 2018. Figure 1 shows the overall counts; Figure 2 provides the percentage of NATO and non-NATO countries with a cyber force. A summary of cyber force structures for both NATO and non-NATO countries is provided in the Appendix for the year 2018, the latest year for which the dataset has been updated.

FIGURE 1: THE TOTAL GROWTH OF CYBER FORCES IN AND OUTSIDE NATO
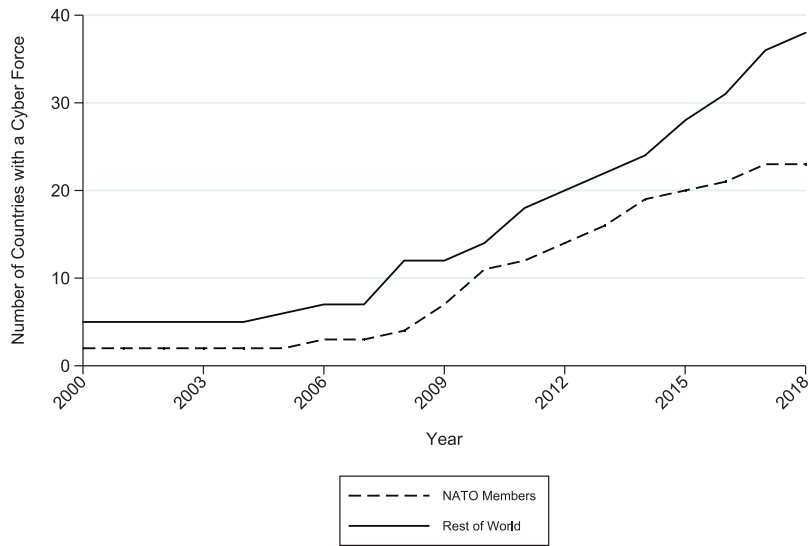
Source: Dataset on Cyber Force Structures

FIGURE 2: THE PERCENTAGE OF STATES WITH A CYBER FORCE IN AND OUTSIDE NATO
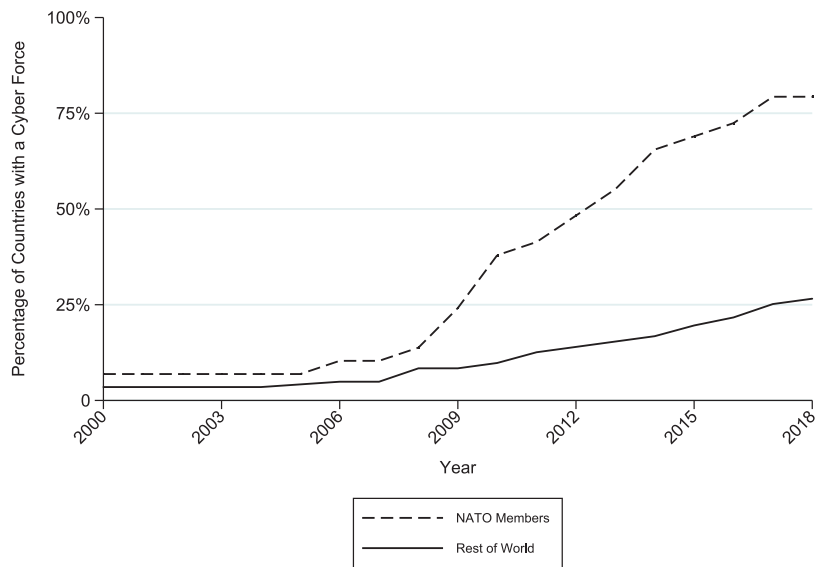
Source: Dataset on Cyber Force Structures

In 2000–2004, only seven countries maintained cyber forces: the United States, Russia, China, Israel, North Korea, Greece, and Thailand each had cyber forces prior to 2000. Worth noting is the consistent increase in cyber forces post-2007. In 2007, there were a total of 10 cyber forces: three in NATO and seven outside NATO. By 2018, there were 61 cyber forces (35.5 percent of militaries) across the world—an average global growth rate of 2.7 percent (four to five new cyber forces) per year since 2007. NATO members accounted for 23 of these 61 forces (Blessing 2020a).

As Figure 1 indicates, non-NATO cyber forces outnumbered NATO-member cyber forces between 2000 and 2018. This trend will inevitably continue, as the number of non-NATO countries is far greater than the number of NATO members. However, Figure 2 provides additional context: cyber forces have emerged at a much faster rate among NATO members than among non-NATO countries. This is particularly clear from 2008 to 2018. Less than 25 percent of NATO countries had a cyber force in 2008. By 2018, nearly 80 percent of NATO countries had developed a cyber force. Conversely, not until after 2017 were there cyber forces in more than 25 percent of non-NATO countries. Thus NATO members have created cyber forces more quickly than the rest of the world; the data suggest that the alliance may be playing a facilitating role. Although the growth of cyber forces in non-NATO states will eventually outpace that of the remaining NATO members over time, NATO countries have led the way in developing military organizations for cyberspace.

Figure 3 shows the global growth of the branch, service, and joint models over time. Significantly, as the number of cyber forces has increased, so has the variation in organizational model. Until 2008, roughly 75 percent of cyber forces utilized the branch model; by 2018, approximately 55 percent of cyber forces used the branch model (a 20 percent drop). While the utilization of the joint model has grown over time, it only accounts for just over 25 percent of the variation by 2018. What Figure 3 indicates is that, although most cyber forces have been structured according to a branch model, the relative prevalence of the branch model has decreased over time. This increasing variation over time runs counter to expectations regarding the emergence of a dominant organizational model.

FIGURE 3: THE WORLDWIDE GROWTH OF CYBER FORCES BY ORGANIZATIONAL MODEL
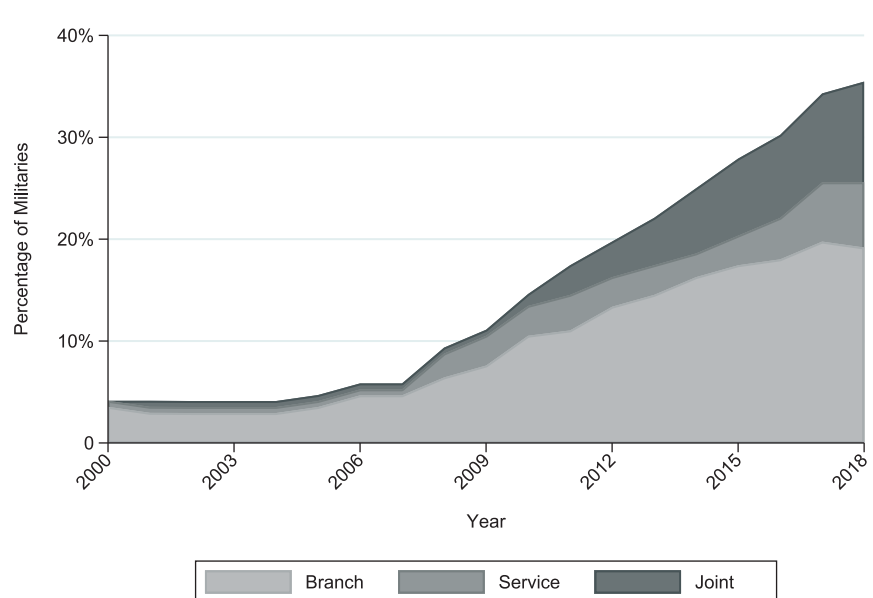
Source: Dataset on Cyber Force Structures

Figure 4 breaks down the variation in organizational model for the cyber forces of NATO members and non-NATO states (2000–2018), while Figure 5 shows the proportion of subordinated, sub-unified, and unified commands from 2009 to 2018 for cyber forces in and outside NATO.

As with Figure 3, the distributions of organizational models in Figure 4 indicate increasing variation over time across cyber forces in both NATO member states (left) and non-NATO states (right). While the branch model has accounted for most cyber force structures, its usage has declined in both groups over time (although somewhat more consistently in non-NATO states). Notably, non-NATO states have opted for the service model at a higher rate than NATO members, while NATO members have used the joint model more extensively than the service model. However, as of 2018, there was no dominant organizational model.

There are several takeaways from Figure 5. First, sub-unified commands only emerge in 2010; the three to appear in 2010 were South Korea's Cyber Command (sub-unified branch), U.S. Cyber Command (sub-unified joint), and Iran's Cyber Defense Command (sub-unified joint). Second, unified commands appear only after 2012 (Norway's Cyber Defense Force, a unified branch, was the first). Third, subordinated commands have been the most prevalent command in non-NATO states. However, by 2018 only half of non-NATO cyber forces were a subordinated command; less than 20 percent had implemented a unified command. Conversely, nearly 40 percent of NATO cyber forces were a unified command by 2018, and less than 20 percent were a subordinated command. On average, a greater proportion of NATO member cyber forces were able to develop into sub-unified and unified commands than non-NATO cyber forces.

FIGURE 4: MODEL DISTRIBUTION ACROSS CYBER FORCES

Source: Dataset on Cyber Force Structures
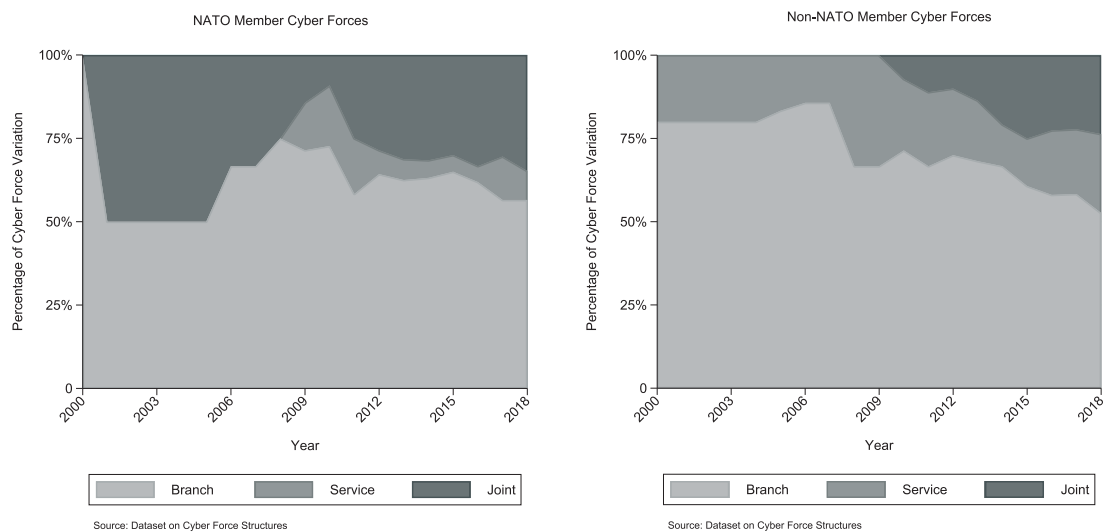
FIGURE 5: COMMAND DEVELOPMENTS IN AND OUTSIDE NATO

Source: Dataset on Cyber Force Structures

Source: Dataset on Cyber Force Structures
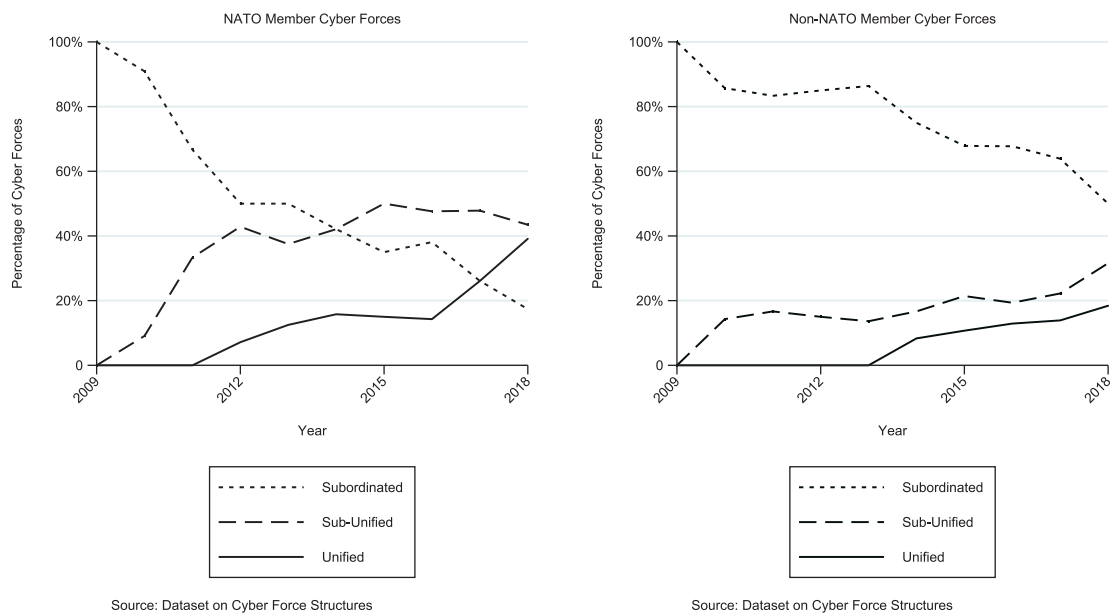
Collectively, Figures 2 through 4 indicate that a subordinated branch has been the most prevalent cyber force structure for both NATO members and non-NATO states. Concluding that this is the predominant force structure, however, is misleading. Each of these figures shows increasing variation over time in both organizational model and scale of command. As new cyber forces were created and existing ones elevated within militaries, there was a decline in the use of the subordinated branch structure relative to other force structures. With the move away from subordinated branches towards unified force structures, can unified cyber forces provide insight into an emerging dominant force structure?

Even across unified cyber forces, the data show variation. Table II looks at all unified cyber force structures in 2018. Across NATO member states as well as rest of the world, unified joint force structures (nine total) were only slightly favored over unified branch (five total) or unified service (two total) arrangements. With only 16 total cyber forces at the unified command level, the unified joint cyber force structure is by no means the predominant paradigm. Evidence thus suggests that, instead of conforming to a single cyber force structure, states have tailored the creation and implementation of cyber forces to their own respective circumstances.

TABLE II: UNIFIED CYBER FORCE STRUCTURES, 2018

	NATO Members	Rest of World	Total
Unified Branch	3	2	5
Unified Service	1	1	2
Unified Joint	5	4	9
Total	9	7	16

Given the variation in cyber force structure across the globe—and between NATO-member states and non-NATO states—what factors can explain force structure choices? Arguably, joint models require greater resource levels and redundant capabilities across combat services than do service or branch models. Likewise, scales of command are likely to be influenced by military spending levels, the size of the workforce, and strategic development. This could be one reason why the joint model and unified commands are somewhat more prevalent across NATO countries: the world's largest economies are disproportionally represented in NATO compared to the rest of the world (World Bank 2018). Although outside the scope of this article, examining the relative influence of these factors on force structure selection and change over time represents fertile ground for future research.

5. IMPLICATIONS FOR NATO

While the force structure data presented above shed light on the cyber force initiatives across NATO's member countries, the individual force structure decisions of states also affect how NATO itself approaches the cyber domain. This paper's findings carry three main implications for NATO.

First and foremost, the rapid increase in the number of NATO members with cyber forces necessitates the development of robust frameworks for integrating sovereign cyber effects into NATO operations (North Atlantic Treaty Organization 2018). The goal of these efforts should be for the alliance to achieve greater effectiveness in cyberspace, particularly as the Cyber Operations Centre relies on personnel from member states with varying capability levels. Additionally, the alliance must start to grapple with the implications of out-of-network operations conducted by members on other allies' networks (Smeets 2019b). At the same time, NATO must account for the inevitable increase in military footprints in cyberspace emerging outside the alliance. The alliance has been at the forefront in setting the international agenda for

cyber issues (Brent 2019). As more states develop cyber forces and existing forces become more mature, NATO and its members are presented with new opportunities to collaborate with non-NATO states.

In this regard, one fruitful way forward for the alliance is to strengthen existing partnerships with non-NATO states and entities. Similar to the 2016 Joint Declaration on NATO-EU Cooperation, the alliance should look to build on its relationships cultivated through the Partnership Interoperability Initiative launched in 2014 (North Atlantic Treaty Organization 2020). More specifically, the alliance could benefit from new initiatives with Enhanced Opportunity Partners like Sweden and Finland, both of which have participated in NATO cyber defense exercises; the latter has also signed the 2017 Political Framework Arrangement on cyber defense cooperation with NATO. Additionally, the alliance could seek to create stronger ties with Australia, a country that has been explicit about its pursuit of offensive capabilities and norm-building in cyberspace (Uren 2018). Other Interoperability Platform Partners like Austria, Japan, South Korea, and Switzerland also offer opportunities to build bridges with established cyber forces. With a broader set of partners, NATO can seek to exchange concepts and develop best practices, test these in exercises, and draw lessons for capability development.

Second, this paper's conceptual framework is important for NATO's net assessment efforts. Force structure is a key aspect of net assessment; however, many elements of traditional force structure become problematic when applied to cyber forces. Several examples illuminate the necessity of this paper's typology for net assessment. Unlike unit functions in other domains,[7] operational functions in the cyber domain can be nearly indistinguishable. Both attacking a network and defending one's own can rely on intrusions into an adversary's networks for intelligence collection. Network exploitation, defense, and attack also use similar tools and techniques (Buchanan 2017, 15–96).

Moreover, instead of tangible weapons systems (missiles, tanks, submarines, etc.) that have multiple-use ability and are quantifiable, "cyberweapons" are comprised of largely digital, transitory elements that have only a temporary ability to access and attack computer networks and systems (Smeets 2018). Capabilities also rapidly diffuse to others: after detecting and patching vulnerabilities after an attack, adversaries can modify and redeploy a capability against the original attacker (Buchanan 2016). Finally, while conventional personnel can be assessed according to the number of direct and indirect military personnel per unit, cyber force personnel complicate net assessments. While total personnel can be quantified, there is no clear distinction between direct "combat" and indirect "support" personnel in the cyber domain. Indirect roles—like signals intelligence—are at the heart of operations for cyber forces' direct personnel.

---

7   Such as armored combat and infantry in the land domain; aircraft carriers and amphibious ships in the maritime domain; bombers and airlift in the air domain; and special operations across domains.

For these reasons, this paper's typology provides the alliance an alternative way to assess force structure; this is particularly important should NATO establish an Office of Net Assessment, as recommended by the NATO 2030 Reflection Group (NATO Reflection Group 2020, 24).

Third, and more broadly, this paper highlights the need for NATO to develop a strategic political framework for coordinating military cyber defense for the alliance and its members. The 2010 Strategic Concept gave relatively little attention to military cyber defense; in fact, the document only uses the word "cyber" five times (North Atlantic Treaty Organization 2010). The data presented in this article indicate that conditions are ripe for integrating cyber defense into the alliance's future strategic concepts. NATO's 2016 recognition of cyberspace as an operational domain, the Cyber Defense Pledge among members, and the establishment of the Cyber Operations Centre have been important milestones. However, the alliance should better define how cyberspace relates to existing core tasks of collective defense, crisis management, and cooperative security. Integrating cyber capabilities into collective defense efforts looms particularly large. For example, the disparity in force structures among members highlights the need to develop a strategy for multi-domain operations, as different force structures are likely to emphasize different operational experiences and approaches to combining cyber capabilities with more traditional ones.

6. CONCLUSION

This paper has offered a comparative perspective of cyber forces and has introduced a new database that catalogues cyber forces from 2000 to 2018. It has also presented a new framework—based on both organizational model (branch, service, or joint model) and the scale of command (subordinated, sub-unified, or unified)—to identify nine unique cyber force structures.

Empirical analysis using this new dataset shows that in 2000, only seven UN-member states possessed cyber forces; 61 UN-member states had created a cyber force by 2018. The data portray consistent growth in the number of cyber forces worldwide; concomitantly, there has been increasing variation in cyber force structure over time. Contrary to conventional expectations, analysis shows that no dominant trends have emerged across either NATO member states or non-NATO states.

Future research can expand on this paper's analysis in several ways. This article did not address why a specific organizational model was chosen for cyber forces; future work can investigate the factors behind model selection for cyber forces. Additionally, future work can explore the facilitators and barriers behind decisions to change force

structure. In this regard, case study research and process tracing political decision-making offer a fruitful way forward. Finally, research can assess how cyber forces change over the course of implementation efforts within militaries. This paper has offered only a static view of the development of cyber forces; a more dynamic view of cyber forces over time is necessary to understand the changing ways in which militaries approach the cyber domain.

REFERENCES

Albuquerque, Adriana Lins de, and Jakob Hedenskog. 2016. "Moldova: A Defence Sector Reform Assessment." FOI-R--4350--SE. Stockholm, Sweden: Swedish Defence Research Agency. https://www.foi.se/rest-api/report/FOI-R--4350--SE.

Applegate, Scott D. 2012. "Leveraging Cyber Militias as a Force Multiplier in Cyber Operations." Fairfax, VA: Center for Secure Information Systems, George Mason University.

Augier, Mie, Thorbjorn Knudsen, and Robert M. McNab. 2015. "Advancing the Field of Organizations through the Study of Military Organizations." *Industrial and Corporate Change* 23 (6): 1417–44.

Baezner, Marie. 2020. "Study on the Use of Reserve Forces in Military Cybersecurity: A Comparative Study of Selected Countries." Zurich, Switzerland: Center for Security Studies, ETH Zurich. https://doi.org/10.3929/ethz-b-000413590.

Blessing, Jason. 2020a. "The Dataset on Cyber Force Structures." Unpublished raw data.

———. 2020b. "The Diffusion of Cyber Forces: Military Innovation and the Dynamic Implementation of Cyber Force Structure." Dissertation, Syracuse University, Syracuse, NY.

Brangetto, Pascal. 2015. "National Cyber Security Organisation: France." National Cyber Security Organisation. Tallinn, Estonia: NATO CCD COE. https://ccdcoe.org/library/publications/national-cyber-security-organisation-france/.

Brenner, Susan W., and Leo L. Clarke. 2011. "Conscription and Cyber Conflict: Legal Issues." In *2011 3rd International Conference on Cyber Conflict*, edited by C. Czosseck, E. Tyugu, and T. Wingfield, 1–12. Tallinn, Estonia: CCD COE Publications.

Brent, Laura. 2019. "NATO's Role in Cyberspace." *NATO Review* (blog). February 12, 2019. https://www.nato.int/docu/review/articles/2019/02/12/natos-role-in-cyberspace/index.html.

Brooks, Risa. 2006. "An Autocracy at War: Explaining Military Effectiveness, 1967 and 1973." *Security Studies* 15 (3): 396–430.

Buchanan, Benjamin. 2016. "The Life Cycles of Cyber Threats." *Survival* 58 (1): 39–58.

———. 2017. *The Cybersecurity Dilemma: Hacking, Trust, and Fear between Nations*. Oxford: Oxford University Press.

Cendoya, Alexander. 2016. "National Cyber Security Organisation: Spain." National Cyber Security Organisation. Tallinn, Estonia: NATO CCD COE. https://ccdcoe.org/library/publications/national-cyber-security-organisation-spain/.

Congressional Budget Office. 2016. "The U.S. Military's Force Structure: A Primer." Washington, D.C.: Congress of the United States. https://apps.dtic.mil/dtic/tr/fulltext/u2/1014153.pdf.

Costello, John, and Joe McReynolds. 2018. "China's Strategic Support Force: A Force for a New Era." 13. China Strategic Perspectives. Washington, D.C.: Center for the Study of Chinese Military Affairs, Institute for National Strategic Studies, National Defense University. https://ndupress.ndu.edu/Portals/68/Documents/stratperspective/china/china-perspectives_13.pdf.

Curley, Gregg. 2018. "The Provision of Cyber Manpower: Creating a Virtual Reserve." *MCU Journal* 9 (1): 191–217.

Dialogo. 2013. "Colombia Rises to the Cyber Challenge," April 1, 2013. https://dialogo-americas.com/en/articles/colombia-rises-cyber-challenge.

DiMaggio, Paul J., and Walter W. Powell. 1983. "The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields." *American Sociological Review* 48 (2): 147–60.

Directorate of Social Communication of the Joint Command of the Armed Forces of Ecuador. 2015. "Fuerzas Armadas realiza taller para defini Infraestructura critica" [Armed Forces conducts workshop to define Critical Infrastructure]. *Nota Periodistica No. 2015-04-20-01-DIR-C.S.*, April 20, 2015. https://www.ccffaa.mil.ec/2015/04/20/fuerzas-armadas-realiza-taller-para-definir-infraestructura-critica/.

Ertan, A., K. Floyd, P. Pernik, and T. Stevens, eds. 2020. "Cyber Threats and NATO 2030: Horizon Scanning and Analysis." NATO CCD COE Publications. https://ccdcoe.org/uploads/2020/12/Cyber-Threats-and-NATO-2030_Horizon-Scanning-and-Analysis.pdf.

Estonian Defence Forces. n.d. "Cyber Command." http://www.mil.ee/en/landforces/Cyber-Command.

Farrell, Theo. 1996. "Figuring Out Fighting Organisations: The New Organisational Analysis in Strategic Studies." *Journal of Strategic Studies* 19 (1): 122–35.

———. 2010. "Improving in War: Military Adaptation and the British in Helman Province, Afghanistan, 2006–2009." *Journal of Strategic Studies* 33 (4): 567–594.

Felongco, Gilbert P. 2016. "Philippine Armed Forces Build Up Capability to Fight in Cyberspace." *Gulf News*, November 23, 2016. https://gulfnews.com/world/asia/philippines/philippine-armed-forces-build-up-capability-to-fight-in-cyberspace-1.1934044.

Gelzis, Gederts. 2014. "Latvia Launches Cyber Defence Unit to Beef Up Online Security." *Deutsche Welle*, March 4, 2014. https://www.dw.com/en/latvia-launches-cyber-defence-unit-to-beef-up-online-security/a-17471936.

Gorwa, Robert, and Max Smeets. 2019. "Cyber Conflict in Political Science: A Review of Methods and Literature." Working Paper Prepared for the 2019 ISA Annual Convention, Toronto, Canada,1–24. URL: 10.31235/osf.io/fc6sg.

Gramaglia, Matteo, Emmet Tuohy, and Piret Pernik. 2013. "Military Cyber Defense Structures of NATO Members: An Overview." Background Paper. Tallinn, Estonia: International Centre for Defence and Security (RKK/ICDS). https://icds.ee/wp-content/uploads/2013/Military%20Cyber%20Defense%20Structures%20of%20NATO%20Members%20-%20An%20Overview.pdf.

Hasik, James. 2016. "Mimetic and Normative Isomorphism in the Establishment and Maintenance of Independent Air Forces." *Defense & Security Analysis* 32 (3): 256–263.

Healey, Jason, ed. 2013. *A Fierce Domain: Conflict in Cyberspace, 1986 to 2012*. Arlington, VA: Cyber Conflict Studies Association.

International Institute for Strategic Studies. 2013. "Europe." In *The Military Balance* 113: 89–198.

———. 2019. "Asia." In *The Military Balance* 119: 222–319.

Italian Ministry of Defence. 2018. "Il Sottosegratario Tofalo visita il Comando C4 Difesa e il CIOC" [Undersecretary Tofalo visits the C4 Defense Command and the CIOC], August 1, 2018. https://www. difesa.it/Primo_Piano/Pagine/Il-Sottosegretario-Tofalo-visita-il-Comando-C4-Difesa-e-il-CIOC.aspx.

Kaska, Kadri. 2015. "National Cyber Security Organisation: The Netherlands." National Cyber Security Organisation. Tallinn, Estonia: NATO CCD COE. https://ccdcoe.org/library/publications/national-cyber-security-organisation-the-netherlandskadri-kaskaactive-passive-cyber-defence-law-national-frameworks-policy-strategy-the-netherlands/.

Keck, Zachary. 2014. "South Korea Seeks Offensive Cyber Capabilities." *The Diplomat*, October 11, 2014. https://thediplomat.com/2014/10/south-korea-seeks-offensive-cyber-capabilites/.

Kong, Ji Young, Jong In Lim, and Kyoung Gon Kim. 2019. "The All-Purpose Sword: North Korea's Cyber Operations and Strategies." In *2019 11th International Conference on Cyber Conflict: Silent Battle*, edited by Tomas Minarik, S. Alatulu, M. Signoretti, I. Tolga, and G. Visky, 143–62. Tallinn, Estonia: CCD COE Publications.

Lasoen, Kenneth L. 2019. "Belgian Intelligence SIGINT Operations." *International Journal of Intelligence and Counterintelligence* 32 (1): 1–29.

Lewis, James Andrew, and Gotz Neuneck. 2013. "The Cyber Index: International Security Trends and Realities." New York and Geneva: United Nations Institute for Disarmament Research.

Lilly, Bilyana, and Joe Cheravitch. 2020. "The Past, Present, and Future of Russia's Cyber Strategy and Forces." In *20/20 Vision: The Next Decade*, edited by T. Jančárková, L. Lindström, M. Signoretti, I. Tolga, and G. Visky, 129–155. Tallinn, Estonia: NATO CCD COE Publications.

Miller, Drew, Daniel B. Levine, and Stanley A. Horowitz. 2013. "A New Approach to Force-Mix Analysis: A Case Study Comparing Air Force Active and Reserve Forces Conducting Cyber Missions." IDA Paper P-4986. Alexandria, VA: Institute for Defense Analyses.

Ministry of Defense of Norway. 2012. "Cyberforsvaret offisielt etablert i dag" [Cyber Defence Force officially established today]. September 18, 2012. https://www.regjeringen.no/no/dokumentarkiv/stoltenberg-ii/fd/Nyheter-og-pressemeldinger/Nyheter/2012/cyber/id699271/.

Moury, Taciana. 2017. "Brazilian Army Invests in Cyber Defense." *Dialogo*, May 12, 2017. https://dialogo-americas.com/en/articles/brazilian-army-invests-cyber-defense.

NATO Reflection Group. 2020. "NATO 2030: United for a New Era." https://www.nato.int/nato_static_fl2014/assets/pdf/2020/12/pdf/201201-Reflection-Group-Final-Report-Uni.pdf.

Nielsen, S. 2016. "The Role of the U.S. Military in Cyberspace." *Journal of Information Warfare* 15 (2): 27–38.

North Atlantic Treaty Organization. 2010. "Active Engagement, Modern Defense: Strategic Concept for the Defence and Security of the Members of the North Atlantic Treaty Organization." https://www.nato.int/nato_static_fl2014/assets/pdf/pdf_publications/20120214_strategic-concept-2010-eng.pdf.

———. 2017. "NATO Supports Jordan's National Cyber Defence Strategy," July 19, 2017. https://www.nato.int/cps/en/natohq/news_146287.htm.

———. 2018. "Framework Mechanism for the Integration of Sovereign Cyber Effects Provided Voluntarily by Allies into Alliance Operations and Missions." North Atlantic Treaty Organization.

———. 2020. "Partnership Interoperability Initiative," November 3, 2020. https://www.nato.int/cps/en/natohq/topics_132726.htm.

Omonobi-Abuja, Kingsley. 2018. "Nigerian Army's Cyber Warfare Command Begins Operation." *Vanguard*, August 29, 2018. https://www.vanguardngr.com/2018/08/nigerian-armys-cyber-warfare-command-begins-operation/.

Osula, Anna-Maria. 2015a. "National Cyber Security Organisation: Estonia." National Cyber Security Organisation. Tallinn, Estonia: NATO CCD COE. https://ccdcoe.org/library/publications/national-cyber-security-organisation-estonia/.

———. 2015b. "National Cyber Security Organisation: United Kingdom." National Cyber Security Organisation. Tallinn, Estonia: NATO CCD COE. https://ccdcoe.org/library/publications/national-cyber-security-organisation-united-kingdom/.

Pernik, Piret. 2018. "Preparing for Cyber Conflict: Case Studies of Cyber Command." Tallinn, Estonia: International Centre for Defence and Security (RKK/ICDS).

Resende-Santos, Joao. 2007. *Neorealism, States, and the Modern Mass Army*. Cambridge: Cambridge University Press.

Robinson, Neil, Agnieszka Walczak, Sophie-Charlotte Brune, Alain Esterle, and Pablo Rodriguez. 2013. "Stocktaking Study of Military Cyber Defence Capabilities in the European Union (MilCyberCAP): Unclassified Summary." RAND Corporation.

Rosen, Stephen Peter. 1991. *Winning the Next War: Innovation and the Modern Military*. Ithaca and London: Cornell University Press.

Schneider, Jacquelyn. 2019. "Deterrence in and through Cyberspace." In *Cross-Domain Deterrence: Strategy in an Era of Complexity*, edited by Erik Gartzke and John Lindsay, 95–120. Oxford: Oxford University Press.

Seker, Esnar, and Ihsan Burak Tolga. 2018. "National Cyber Security Organisation: Turkey." National Cyber Security Organisation. Tallinn, Estonia: NATO CCD COE. https://ccdcoe.org/library/publications/national-cyber-security-organisation-turkey/.

Smeets, Max. 2018. "A Matter of Time: On the Transitory Nature of Cyberweapons." *Journal of Strategic Studies* 41 (1–2): 6–32.

———. 2019a. "NATO Members' Organizational Path Towards Conducting Offensive Cyber Operations: A Framework for Analysis." In *11th International Conference on Cyber Conflict: Silent Battle*, edited by T. Minarik, S. Alatulu, S. Biondi, M. Signoretti, I. Tolga, and G. Visky, 163–78. Tallinn, Estonia: NATO CCD COE Publications.

———. 2019b. "NATO Allies Need to Come to Terms With Offensive Cyber Operations." *Lawfare* (blog). October 14, 2019. https://www.lawfareblog.com/nato-allies-need-come-terms-offensive-cyber-operations.

Uren, Tom. 2018. "Australia's Offensive Cyber Capability." *The Strategist, Australian Strategic Policy Institute* (blog). April 10, 2018. https://www.aspistrategist.org.au/australias-offensive-cyber-capability/.

U.S. Cyber Command. n.d. "U.S. Cyber Command History." Accessed July 13, 2019. https://www.cybercom.mil/About/History/.

Valeriano, Brandon, and Ryan C. Maness. 2015. *Cyber War Versus Cyber Realities: Cyber Conflict in the International System*. Oxford and New York: Oxford University Press.

Wiener, Craig J. 2016. "Penetrate, Exploit, Disrupt, Destroy: The Rise of Computer Network Operations as a Major Military Innovation." Doctoral dissertation, George Mason University, Fairfax, VA.

World Bank. 2018. "World Development Indicators." Washington, D.C.: World Bank.

APPENDIX: NATO AND NON-NATO CYBER FORCE STRUCTURES, 2018

The data presented below describes cyber force structures for the year 2018, the most recent year for which the Dataset on Cyber Force Structures (DCFS) has been updated. Because the database was originally presented as part of doctoral dissertation work in Blessing (2020b), the first round of data collection efforts, covering the period between January 2000 and December 2018, concluded in 2019. As such, the data below do not reflect the most up-to-date force structures for each country. A second round of data collection and coding, which will update the DCFS for the 2019–2021 period, is currently underway and is scheduled to be completed in early 2022.

The organizational names provided below correspond to official national sources and have been translated into English.

TABLE III: CYBER FORCE STRUCTURE FOR NATO MEMBER STATES, 2018

Country	Organization Name	Organizational Model	Scale of Command
Albania	Defense Intelligence and Security Agency	branch	subordinated
Belgium	Cyber Security Operations Centre	branch	sub-unified
Canada	Directorate of Cybernetics	branch	sub-unified
Croatia	Center for Communications and Information Systems	branch	sub-unified
Czechia	National Cyber Operations Centre	branch	sub-unified
Denmark	Computer Network Operations Unit	joint	subordinated
Estonia	Cyber Command	branch	unified
France	Cyber Defense Command Unit	joint	unified
Germany	Cyber and Information Space Command	service	unified
Greece	Joint Cyber Command	joint	unified
Hungary	Cyber Defense Center	branch	sub-unified
Italy	Joint Command for Cyberspace Operations	joint	sub-unified
Luxembourg	Army Cyber Cell	service	subordinated
Netherlands	Defense Cyber Command	joint	unified
Norway	Cyber Defense Force	branch	unified
Poland	Cyber Operations Centre	branch	sub-unified
Portugal	Cyber Defense Centre	branch	subordinated
Romania	Cyber Defense Command	branch	unified
Slovakia	Cyber Defense Centre	branch	sub-unified
Spain	Joint Cyber Defense Command	joint	unified
Turkey	Turkish Armed Forces Cyber Defense Command	branch	sub-unified
United Kingdom	Joint Forces Cyber and Electromagnetic Group	joint	sub-unified
United States	U.S. Cyber Command	joint	unified

TABLE IV: CYBER FORCE STRUCTURE FOR NON-NATO STATES, 2018

Country	Organization Name	Organizational Model	Scale of Command
Argentina	Joint Cyber Defense Command	joint	unified
Australia	Defense SIGINT and Cyber Command	joint	sub-unified
Austria	Command Support and Cyber Defense Command	branch	unified
Bangladesh	Directorate General of Forces Intelligence	branch	subordinated
Belarus	Army Cyber Units	service	subordinated
Brazil	Cyber Defense Command	service	sub-unified
Chile	Joint Cyber Defense Command	joint	sub-unified
China	People's Liberation Army Strategic Support Force	service	unified
Colombia	Joint Cybersecurity and Cyber Defense Command	joint	sub-unified
Ecuador	Cyber Defense Command	joint	unified
Finland	Cyber Defense Division	branch	sub-unified
India	Defense Information Warfare Agency	branch	subordinated
Indonesia	Cyber Operations Command	branch	sub-unified
Iran	Cyber Defense Command	joint	sub-unified
Ireland	Communications and Information Services Corps	branch	subordinated
Israel	Unit 8200	branch	subordinated
Japan	Cyber Defense Unit	branch	sub-unified
Kazakhstan	Cyber Branch	branch	unified
Malaysia	Cyber Defense Operation Center	branch	subordinated
Mexico	Naval Cybersecurity Center	service	subordinated
Myanmar	Military Security Affairs	branch	subordinated
Nigeria	Cyber Warfare Command	service	sub-unified
North Korea	Unit 121	branch	subordinated
Paraguay	General Directorate of Information Technology and Communication	branch	sub-unified
Peru	Cyber Defense Command	service	sub-unified
Philippines	AFP Signal Corps	service	subordinated
Russia	Main Directorate of the General Staff (GRU)	branch	subordinated
Serbia	Command Information Systems and IT Support Centre	branch	subordinated
Singapore	Cyber Defense Group	branch	subordinated
South Africa	Defense Intelligence Division	branch	subordinated
South Korea	Defense Cyber Command	joint	unified
Sri Lanka	Army Signals Corps 12th Regiment	service	subordinated
Sweden	Military Intelligence and Security Service	branch	subordinated
Switzerland	Electronic Operations Centre	branch	subordinated
Thailand	Army Cyber Center	service	subordinated
Ukraine	Main Directorate of Communication and Information Systems	branch	subordinated
Venezuela	Joint Directorate of Cyber Defense	joint	sub-unified
Vietnam	Cyberspace Operations Command	joint	unified

Windmills of the Mind: Higher-Order Forms of Disinformation in International Politics

James Shires

Assistant Professor, Cybersecurity Governance
Institute of Security and Global Affairs
University of Leiden
The Netherlands
j.shires@fgga.leidenuniv.nl

Abstract: Disinformation – the organised and deliberate circulation of verifiably false information – poses a clear danger to democratic processes and crisis response, including the current coronavirus pandemic. This paper argues for a conceptual step forward in disinformation studies, continuing a trend from the identification of specific pieces of disinformation to the investigation of wider influence campaigns and strategic narrative contestation. However, current work does not conceptually separate first-order forms of disinformation from higher-order forms of disinformation: essentially, the difference between disinformation about political or other events, and disinformation *about disinformation itself*.

This paper argues that this distinction is crucial to understanding the extent and consequences (or lack thereof) of disinformation in international politics. The paper first highlights how political disinformation is often sparked by leaks – the release of secret or confidential information into the public domain. It suggests that disinformation and leaks intersect with conventional cybersecurity threats through the increasingly common phenomenon of hack-and-leak operations. The paper then introduces the concept of higher-order disinformation. This discussion is followed by an empirical example: the case of US intelligence assessments of Russian hack-and-leak operations during the US presidential election campaign in 2016. The paper concludes with offensive and defensive policy implications, arguing that the relevance of second, third, and higher orders of disinformation will only increase as more experienced actors draw on the material, successes, and lessons of previous campaigns.

Keywords: *disinformation, hack-and-leak operations, leaks, Russia, US, narrative*

1. INTRODUCTION

Disinformation is an essentially social problem: in one useful definition, it is not simply misleading communication but communication that has *the central function of misleading* a specific audience.[1] However, it would be a mistake to see disinformation simply as a defect in systems of communication, whether written, verbal, or visual. Such an approach, common in both policy and academic literature on disinformation, draws on a simple "transmission" view of communication. It approaches political communities in an almost cybernetic fashion, focusing on the extent to which accurate information is transferred between different parts of the system.[2]

However, disinformation, along with a broader array of misdirection and deception, is not a secondary add-on to or corruption of pure information flows in an ideal body politic but an integral part of that political community. The community itself would not exist without the rumours, lies, and half-truths that circulate within it.[3] In this view, there is no such thing as "pure" – unbiased, not slanted, non-ideologically committed – communication against which to compare clear examples of disinformation. To continue the biological metaphor, just as bacteria are not an external, negative threat to biological organisms but a central part of their inner constitution, the same applies to societies and disinformation. Consequently, although the theme of this conference is "going viral", viruses – especially in the current pandemic times – are a misleading "organizing metaphor" for disinformation: a better one is bacterial.[4]

This is not a new insight, and most approaches to political science and international relations recognise that questions of truth and falsity cannot be answered without considering broader issues around discursive power and silence, and narrative construction and contest.[5] In studies of disinformation more specifically, this insight has encouraged a trend away from the identification of specific pieces of disinformation, to be countered by education and fact-checking, to the investigation of wider "influence" campaigns. Such campaigns are often identified and investigated along the lines of Advanced Persistent Threat (APT) methodologies in cybersecurity, notably demonstrated in high-profile "takedowns" by large platform companies – a point to which I return below.[6]

---

1   Alexander Lanoszka, "Disinformation in International Politics", *European Journal of International Security* 4, no. 2 (June 2019): 227–248, https://doi.org/10.1017/eis.2019.6.

2   For an example of this approach, see Bruce Schneier and Henry Farrell, "Common-Knowledge Attacks on Democracy" (Berkman Klein Center for Internet and Society, Harvard University, October 2018).

3   Sally Engle Merry, "Rethinking Gossip and Scandal", in *Toward a General Theory of Social Control: Fundamentals*, edited by Donald Black (Orlando and London: Academic Press, 1984), 271–302.

4   Jordan Branch, "What's in a Name? Metaphors and Cybersecurity", *International Organization* 75, no. 1 (2021): 39–70, https://doi.org/10.1017/S002081832000051X.

5   David Campbell, *Writing Security: United States Foreign Policy and the Politics of Identity* (Manchester: Manchester University Press, 1998); Ronald R. Krebs, *Narrative and the Making of US National Security* (Cambridge University Press, 2015).

6   See, e.g., Facebook, "February 2020 Coordinated Inauthentic Behavior Report", *About Facebook* (blog), March 2, 2020, https://about.fb.com/news/2020/03/february-cib-report/.

This trend has reached its most sophisticated and historically aware treatment in the concept of "active measures", the subject of Thomas Rid's recent book of the same name.[7] Active measures are more than disinformation campaigns: they are (usually or mostly covert) bureaucratic efforts to marshal the combination and spread of information (whether true, false, or somewhere in between) for specific strategic ends, persisting – in Rid's treatment of the Russian case – beyond the lifetime of organisations, technological infrastructures, and even entire political regimes. However, even Rid's exemplary work only identifies instances of first-order forms of disinformation but does not conceptually separate first-order forms from higher-order forms: essentially, the difference between disinformation about political or other events, and disinformation *about disinformation itself*.

In this paper, I argue that this distinction helps us to bridge the two approaches to disinformation above: on the one hand, a "transmission" view of communication in which specific pieces of information have verifiably factual or false content, and on the other hand, a recognition that all communication, especially of the political kind, takes place against a backdrop of powerful discursive presuppositions and broader narrative contest. By considering the reflexive quality of individual instances of disinformation, and their references back to and dependence upon prior contested claims, we can progress analytically from the former view to the latter, tracing how fundamental splits in worldview emerge, stacked upon a succession of divergent factual claims as well as different political commitments. Understanding higher orders of disinformation is thus crucial to understanding the extent and consequences (or lack thereof) of disinformation in international politics overall.

The paper is structured as follows. The following section narrows the focus of the paper from disinformation overall to a specific kind of influence operation – hack-and-leak operations – that, due to their intersection with conventional cybersecurity threats, are a key focus of US defence and cyber policy.[8] The third section uses hack-and-leak operations to introduce the concept of higher-order forms of disinformation, meaning that second-order disinformation is leaks about (alleged) hack-and-leak operations, third-order disinformation is leaks about those leaks, and so on. The fourth section applies this largely abstract discussion to the case of US intelligence assessments of Russian influence operations around the 2016 presidential election. The final section concludes, reflecting on both offensive and defensive policy implications of this paper: offensively, the problems in mounting counter-disinformation disinformation operations, and defensively, the limits of relying on content moderation and fact-checking services to police disinformation.

---

7 Thomas Rid, *Active Measures: The Secret History of Disinformation and Political Warfare* (New York: Profile Books, 2020).

8 Stephen G. Fogarty and Bryan N. Sparling, "Enabling the Army in an Era of Information Warfare", *Cyber Defense Review* 5, no. 2 (Summer 2020). See also US Department of Defense, "Summary: Department of Defense Cyber Strategy", Washington, DC, 2018, p. 1.

2. THE MECHANICS OF HACK-AND-LEAK OPERATIONS

Hack-and-leak operations are, as several scholars have argued, the pinnacle of disinformation operations: they combine a compromise of digital networks to obtain information (hack) with the release of that information for strategic effect (leak).[9] This is not a necessary combination: many hacks occur without compromised information ever coming to light, while many leaks occur through more mundane forms of access – although they are no less dependent on the broader communications ecosystem built around the internet.[10] The paradigm example of a hack-and-leak operation is the release of information gained from the Democratic National Committee (DNC) and related entities and people before the 2016 US election, attributed to Russian intelligence agencies (for the leak, specifically, the military Main Intelligence Directorate [GRU]) by subsequent US government investigations and many independent observers.[11] However, we should not let the impact of this incident on academic and policy research on disinformation leave us unable to see the wood for a single large tree: notable state-sponsored hack-and-leak operations have taken place against international sporting bodies (the World Anti-Doping Agency [WADA], the International Federation of Association Football [FIFA]), private entities in the US (Sony Pictures), and in other national contexts (Macronleaks, the 2019 UK election, and the Saudi cables), as well as in situations of destabilisation and conflict (for example, in many instances in Syria).[12]

Before considering the precise relationship between hack-and-leak operations and disinformation, it is instructive to briefly outline the conceptual mechanics of hack-and-leak operations from an analytical perspective, rather than that of the target or perpetrator. Basically, hack-and-leaks, like leaks more generally, function within larger constructions of privacy and/or secrecy.[13] Words and deeds must first be kept private, or at a national level, classified as secret, to then be leaked. The first conceptual building block of a hack-and-leak is thus the protection and limitation of information. Many analyses omit this element, missing how variations in social expectations of secrecy or technological means for achieving it affect the outcome of

---

9   Jaclyn Alexandra Kerr and Herbert Lin, "On Cyber-Enabled Information/Influence Warfare and Manipulation", SSRN, March 13, 2017; James Shires, "Hack-and-Leak Operations: Intrusion and Influence in the Gulf", *Journal of Cyber Policy* 4, no. 2 (2019): 235–256.

10 Ronald J. Deibert, *Reset: Reclaiming the Internet for Civil Society* (Toronto: House of Anansi Press, 2020).

11 Robert S. Mueller, *Report on the Investigation into Russian Interference in the 2016 Presidential Election*, Submitted Pursuant to 28 C.F.R. § 600.8(c) (Washington, DC: US Department of Justice, March 2019).

12 Ben Buchanan, *The Hacker and the State: Cyber Attacks and the New Normal of Geopolitics* (Cambridge, MA: Harvard University Press, 2020); Jean-Baptiste Jeangene Vilmer, "The 'Macron Leaks' Operation: A Post-Mortem" (Atlantic Council and IRSEM, June 2019); James Shires, "Understanding the Tactics behind Hack-and-Leak Operations", *Atlantisch Perspectief* 4 (September 2020); Marie Baezner, "The Use of Cybertools in an Internationalized Civil War Context: Cyber Activities in the Syrian Conflict", CSS Cyber Defense Project (Center for Security Studies, ETH Zurich, October 18, 2017).

13 Sissela Bok, *Secrets: On the Ethics of Concealment and Revelation* (New York and Toronto: Vintage, 1989).

a leak.[14] Given such a private act, the two other key elements of a leak are *access* – an "outsider" gaining access to the private place – and *dissemination* – the spread of that information once obtained (Figure 1). Of course, both are spectrum rather than binary concepts: insider threats highlight the difficulty in access control, while leaked information rarely emerges into the open in a symmetric, equal fashion. Hack-and-leak operations, as a subset of leaks more broadly, can be defined as those involving a particular means of access: offensive cyber capabilities for remote intrusion into digital networks.

FIGURE 1: CONCEPTUAL MODEL OF HACK-AND-LEAK OPERATIONS
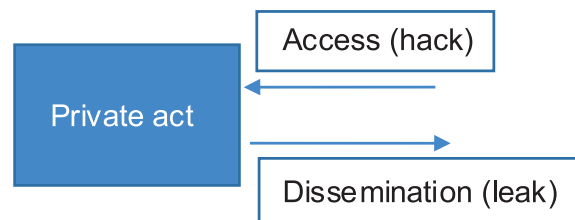
Hack-and-leak operations do not always include disinformation (although they are always "active measures", in Rid's definition of the term). However, it is precisely this expectation of privacy and/or secrecy that makes leaks such powerful vehicles for disinformation through doctoring or altering content. Leaks carry (often erroneous) connotations of franker, more truthful communication, without the many layers of artifice we expect from public political communication. Current scholarship has focused primarily on the *amplifying* relationship between hack-and-leaks and disinformation. Researchers have traced how what François calls "false leaks" spread on social media platforms like Twitter, highlighting how their dissemination through certain hashtags affects their impact.[15] Others have argued that "tainted leaks" of doctored information gained through phishing attacks against journalists and political opponents have been used by the Russian government to "seed mistrust".[16] As Rid demonstrates, these are not new tactics and existed well before the internet.[17] Elsewhere, I have argued that "edge cases" of hack-and-leaks, where almost all the released information is doctored – such as the cyber operation against the Qatar News Agency in 2017 – highlight the shifting boundaries between leaked and manufactured information.[18] Furthermore, the act that is the subject of a leak does not have to be documentary in form: the Shadow Brokers leaks highlight how offensive cyber capabilities can themselves be

---

14   David Pozen, "The Leaky Leviathan: Why the Government Condemns and Condones Unlawful Disclosures of Information", *Harvard Law Review* 127 (February 25, 2013): 512–635.

15   Presentation by Camille François of Graphika at CyberWarCon, Washington, DC, November 2018.

16   Adam Hulcoop et al., "Tainted Leaks: Disinformation and Phishing With a Russian Nexus", *Citizen Lab*, May 25, 2017.

17   Rid, *Active Measures*.

18   James Shires, "The Cyber Operation against Qatar News Agency", in *The 2017 Gulf Crisis: An Interdisciplinary Approach*, edited by Mahjoob Zweiri, M. Mizanur Rahman, and A. Kamal (Berlin and Heidelberg: Springer Nature, 2020).

the subject of (alleged) hack-and-leaks, introducing an entirely new level of damage from their release.[19]

This growing body of scholarship demonstrates how disinformation – be it doctoring, falsifying, forging, or tainting – changes the mechanics of hack-and-leak operations above, in terms of both access and dissemination. If a claimed hack-and-leak is in fact a disinformation operation, then no access to a private or secret space is required. Of course, successful tainting requires raw material, and successful forgeries are usually based on close knowledge of genuine documents, so good access is likely to increase the impact of a disinformation operation based on a "leak" – but it is not necessary. In terms of dissemination, the problem is no longer how to identify relevant information on the target networks and extract it undetected, but how to muddy the sourcing so it *appears* to the eventual audience that a hack-and-leak was a plausible originating point. A good example of such vague genesis is the appearance of controversial documents about National Health Service (NHS) funding shortly before the 2019 UK general election. They first appeared on Reddit and took a while to catch the attention of the media before ending up in the hands of the opposition leader, Jeremy Corbyn, in a national televised debate.[20]

Overall, hack-and-leak operations can be a potentially effective but highly complex vehicle for disinformation. At their most effective, they act as the "simulation of scandal", combining genuine leaked information with difficult-to-detect nuggets of disinformation to embarrass or discredit a target.[21] Such operations may remain undetected or misdescribed for years, and it is likely that the empirical record of hack-and-leak operations only captures a small percentage of the overall cases. But their complexity means that they have several potential pitfalls, not least the law of diminishing returns: frequent scandals mean that audiences may be inured to later leaks, especially if manipulation is commonplace enough that people no longer give greater credence to leaked material. Furthermore, as I have argued elsewhere, hack-and-leak operations often backfire, because media attention and cyber "hype" mean that hacks are as newsworthy as leaked content, if not more so – especially when state-sponsored.[22] However, despite this complexity of effect, the basic mechanics of hack-and-leak operations – access to and dissemination of a private act – are relatively simple. This, in addition to their inclusion in US and other policy priorities, makes them a good focal point for the introduction of higher orders of disinformation in the following section.

19   Buchanan, *The Hacker and the State*.
20   Ben Nimmo et al., "Secondary Infektion", *Graphika*, June 2020.
21   James Shires, "The Simulation of Scandal: Hack-and-Leak Operations, the Gulf States, and U.S. Politics", *Texas National Security Review*, August 2020.
22   Shires, "The Simulation of Scandal".

3. HIGHER-ORDER FORMS OF DISINFORMATION

The concept of higher orders, in abstract terms, is relatively straightforward. For any x, a higher order x is reflexive: it is the application of x to x itself (second order), or the application of x to the application of x to x (third order), and so on. The concept has been used across philosophy and the social sciences, being deployed in the context of everything from conscious awareness (a mental state about a mental state) to the modelling of rational interaction in economics and political science (from x's beliefs about y, to y's beliefs about x's beliefs about y, and so on). Even in these examples, the power of the concept of higher orders should be apparent: it can account for the transition from a simple, single-level phenomenon, to multi-level, complex phenomena, without invoking more and more different types of entities or concepts: the reflexive repetition of a single concept is sufficient to explain the difference in complexity.

Disinformation creates a dilemma that seems – on the face of it – to call for a higher-order conceptual architecture. On the one hand, focusing on the "verifiably false" nature of specific claims leads quickly onto thorny ground.[23] For example, the EU External Action Service (EEAS), an EU agency founded in 2010, has an East StratCom Task Force, established in 2015, which seeks to "increase public awareness and understanding of the Kremlin's disinformation operations".[24] To do this, the EEAS runs a website, *EUvsDisinfo* (euvsdisinfo.eu), with a well-populated "disinfo database" of specific pieces of disinformation archived from media websites in multiple languages, with date, target audience, and other key characteristics. Each piece includes a summary and a "disproof", a body of text that contradicts or debunks the claims made by the disinformation piece. However, in many cases the "disproof" is not exactly that, because the piece of disinformation itself was not precise enough to be debunked. Instead, the "disproof" offers a contrasting narrative, drawing on wider geopolitical statements that, crucially, do not represent a shared ground of agreement (for example, between pro- and anti-Russian government positions). This sustained and careful project, focusing on specific pieces of disinformation, runs aground because it is easily drawn into wider contests over frames and narratives.

On the other hand, as highlighted in the introduction, many recent analyses do not interrogate the "verifiably false" nature of specific claims[25] but instead reorient the debate using terms such as "influence campaigns" or some platform companies' preferred term of "Coordinated Inauthentic Behavior" (CIB). The label of "influence campaigns" echoes the cybersecurity industry's shift away from solely detecting specific cybersecurity incidents or events (analogous to a particular instance of

---

23   European Commission, "A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation" (Luxembourg: European Commission Directorate-General for Communication Networks, Content and Technology, March 2018).

24   EUvsDisinfo, https://euvsdisinfo.eu/about/

25   European Commission, "A Multi-Dimensional Approach to Disinformation".

disinformation) to connecting such incidents together as intrusion campaigns according to common tactics, techniques, and procedures (TTPs) and broader strategic objectives. It is more than an echo, in fact, as cybersecurity professional experience, commercial structures, and even specific APT labels can thus be transferred from intrusion campaigns to the problem of disinformation. Rid's concept of active measures represents the apex of this trend, focusing on the strategic and bureaucratic practices and ideologies underpinning a wide range of campaigns. These concepts are each significantly different, but they all operate at a far more sophisticated level than approaches seeking simply to "disprove" disinformation.

The question, then, is: how can we connect these two approaches to disinformation? I suggest that we can understand how disinformation expands into wider differences in frame and narrative using the concept of higher orders introduced above. Such an analysis begins by identifying key informational nodes that fracture audience perspectives, perceived by some as central factual elements of their overall worldview, and as disinformation by others.[26] Such nodes are the basis for further contentious claims, which revolve around the credibility of earlier nodes. These subsequent claims are, for those who disagree with that interpretation of the informational node, a second-order form of disinformation: disinformation *about disinformation*. These claims in turn invite further claims: third-order or higher forms of disinformation.

The case study in the following section applies this approach to a specific case study; before doing so, I illustrate the approach in more detail using the framework of hack-and-leak operations introduced above. A higher-order treatment of hack-and-leak operations would use only the concepts identified in the previous section (a private or secret act, and access to and dissemination of that act). More specifically, to explain how the hack-and-leak itself becomes the subject of media attention, we can see the hack-and-leak as a second private act, encompassing the original private act (the subject of the hack-and-leak) as well as the access to and dissemination of information. Consequently, this second private act (the whole hack-and-leak) can itself be subject to access (discovering the hack) and dissemination (informing the media that the original scandal was the result of a hack-and-leak). This reflexive step is illustrated in Figure 2.

---

26   This is always a *further* fracturing: there is no single original audience and no cohesive public sphere prior to such disagreements.
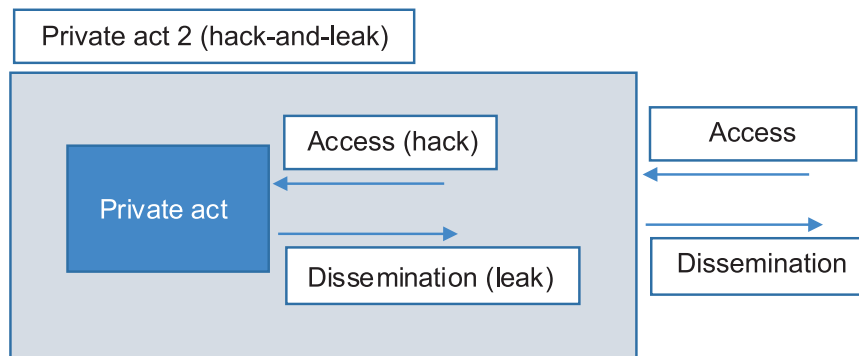
This reflexive application of the same concepts is useful because it explains more complex cases without resorting to a larger conceptual architecture. In cases where the hack becomes as newsworthy a story as the leak (for example, in the case of Russian intrusion into the DNC in 2016), access to the hack-and-leak operation – through Crowdstrike's technical analysis, the Mueller investigation, and many other means – and its dissemination – the Mueller report, congressional testimony, countless media articles, and many other publications – have turned the hack-and-leak into a private act to be revealed to the public in just the way that the original private act (confidential emails and documents) were revealed to the public by the GRU via Wikileaks. Furthermore, this is only the first step in the application of higher-order concepts: as illustrated below, third- and subsequent-order versions quickly emerge.

4. "RUSSIAN INTERFERENCE" VS "RUSSIA HOAX"

This section examines one aspect of the most high-profile example of an influence campaign in recent history: Russian activities relating to the 2016 US presidential election. As noted above, these activities were reported to include – but were not limited to – a hack-and-leak operation against the DNC and related entities. This hack-and-leak operation acted as a key informational node, morphing into two far broader narratives in US politics. One was an anti-Trump narrative of "Russian interference", taking forensic evidence around the DNC compromise and the subsequent Mueller investigation at face value. The other was a pro-Trump narrative of a "Russia hoax", propagated by President Donald J. Trump himself, his family and close associates, right-wing media outlets, and social media commentators.[27] The "Russia hoax" narrative claims that the DNC hack-and-leak operation and wider claims of links between the Trump campaign and the Russian government were part of a deliberate plan to sabotage the Trump campaign and then the presidency itself.

---

[27]   I use pro- and anti-Trump as the most accurate way of designating US political divisions during the 2016–2020 term, rather than Republican/Democrat or left/right-wing.

This case exemplifies the disinformation dilemma I identified above: the movement from specific document leaks to their embedding in larger narratives and frames. It should be stressed that there are long-term reasons for this split in US political worldviews (not least the decade-long evolution of the right-wing media ecosystem),[28] and the key informational node of the hack-and-leak operation fractured these perspectives *further* rather than beginning the process.[29] Nonetheless, its divisive nature and subsequent policy impact make it a crucial case for the conceptual framework of higher-order forms of disinformation introduced above. To focus this brief account, I centre the following discussion on the declassification of documents relating to Russian activities and the 2016 election that occurred at the end of September 2020, ordered by then-Director of National Intelligence (DNI) John Radcliffe. Radcliffe is a prominent Republican and was a member of Congress until his appointment as DNI in May 2020 by President Trump.

This declassification is important for three reasons. The first was its timing: the declassification occurred at a key point in President Trump's run for re-election, and many commentators claimed it was designed specifically to influence the 2020 campaign. Second, the declassification was itself a leak, insofar as it generated significant controversy within former and current members of the intelligence community about whether it conformed to standard practices of declassification, or even specific regulations.[30] Third, the declassified documents connect individual reports from 2016 to the overall split in narratives above, with both sides claiming that the declassified documents support the "Russian interference" and the "Russia hoax" narrative, respectively.[31]

On 29 September 2020, DNI Radcliffe declassified three US intelligence documents, the first two of which were released by Fox News.[32] The first document contained handwritten notes by then-CIA Director John Brennan from a meeting with President Obama in late July 2016, concerning "alleged approval by Hillary Clinton on 28 July of a proposal from one of her foreign policy advisers to vilify Donald Trump by stirring up a scandal claiming interference by the Russian security service".[33] The second document was a CIA memo to the FBI on 7 September 2016, providing "examples of information the Crossfire Hurricane [investigation into Russia links to the Trump

---

28   Yochai Benkler, Robert Faris, and Hal Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (New York: Oxford University Press, 2018).

29   Kathleen Hall Jamieson, *Cyberwar: How Russian Hackers and Trolls Helped Elect a President: What We Don't, Can't, and Do Know* (New York: OUP USA, 2018).

30   Brian Greer, "John Ratcliffe's Dangerous Declassification Game", *Lawfare*, October 7, 2020, https://www.lawfareblog.com/john-ratcliffes-dangerous-declassification-game.

31   Andrew Desiderio and Daniel Lippman, "Intel Chief Releases Russian Disinfo on Hillary Clinton That Was Rejected by Bipartisan Senate Panel", *Politico*, September 29, 2020, https://www.politico.com/news/2020/09/29/john-ratcliffe-hillary-clinton-russia-423022.

32   Brooke Singman, "DNI Declassifies Brennan Notes, CIA Memo on Hillary Clinton 'Stirring up' Scandal between Trump, Russia", Fox News, October 6, 2020, https://www.foxnews.com/politics/dni-brennan-notes-cia-memo-clinton.

33   Ibid.

campaign] fusion cell has gleaned", including "an exchange [redacted] discussing US presidential candidate Hillary Clinton's approval of a plan concerning US presidential candidate Donald Trump and Russian hackers hampering US elections as a means of distracting the public from her use of a private email server".[34] The third document (not released directly) stated that in late July 2016, US agencies "obtained insight into Russian intelligence analysis alleging" that Clinton "approved a campaign plan to stir up a scandal" against Trump "by tying him to Putin and the Russians' hacking of the DNC".[35]

The media and political response to these documents in the US was extremely polarised, across both traditional and social media. Pro-Trump observers took this declassification in the way it was likely intended by the DNI, seeing it as evidence that the US intelligence community knew of improper practices by the Clinton campaign and yet did not follow them up, thus falling for what these observers saw as the "Russia hoax".[36] By contrast, anti-Trump observers largely focused on the role of Russian intelligence analysis as the source of the alleged Clinton plans in these documents, highlighting the declassified sentence that "the IC [intelligence community] does not know the accuracy of this allegation or the extent to which Russian intelligence analysis may reflect exaggeration or fabrication" as a basis for claiming that the declassification was based on "Russian disinformation".[37] Some anti-Trump commentators went further, claiming not only that Russian disinformation was the source of the documents but that the declassification was therefore itself a form of disinformation, as an inappropriate declassification (i.e., a leak) based on false information and designed to mislead.[38] The declassification event clearly resists a neat analytical interpretation, with almost any treatment likely to lean towards one or the other of the two broader narratives of "Russian interference" or "Russia hoax".[39]

The conceptual model of higher-order disinformation can help analysts trace how these two broader narratives relate to the specific declassified documents and the hack-and-leak operation that is their subject, connecting the two levels of analysis identified earlier. Unlike the second-order model presented in the previous section, this case exhibits at least five orders of reflexivity: (5) the DNI's declassification

---

34 Ibid.
35 Ibid.
36 Jerry Dunleavy, "Obama Was Briefed on Unverified Russian Report Claiming Clinton Approved Plan to Tie Trump to Putin and DNC Hack", *Washington Examiner*, September 29, 2020, https://www.washingtonexaminer.com/news/obama-was-briefed-on-unverified-russian-report-claiming-clinton-approved-plan-to-tie-trump-to-putin-and-dnc-hack.
37 Sonam Sheth, "Trump's Spy Chief Just Released 'Russian Disinformation' against Hillary Clinton that He Acknowledged May Be Fabricated", *Business Insider*, September 30, 2020, https://www.businessinsider.in/politics/world/news/trumps-spy-chief-just-released-russian-disinformation-against-hillary-clinton-that-he-acknowledged-may-be-fabricated/articleshow/78396299.cms.
38 Zachary Cohen and Alex Marquardt, "Former CIA Director Accuses Intel Chief of Selectively Declassifying Documents to Help Trump", CNN, October 7, 2020, https://www.cnn.com/2020/10/06/politics/brennan-ratcliffe-declassifying-intelligence-clinton-russia/index.html.
39 This includes the analysis here, which, one reviewer noted, could be construed as "an attack on right-wing politics".

or dissemination of (4) US intelligence community documents about (3) Russian intelligence analysis about (2) an alleged plan by the Clinton campaign to tie together Trump and Russia in the (1) hack-and-leak of documents from the DNC. These are orders of reflexivity, rather than separate events, because they each revolve around the same claims, piling extra layers of interpretation on at each stage. Crucially, each higher order introduces further potential for disinformation, as the key elements of access and dissemination are questioned at each stage: first, for the alleged plan described by Russian intelligence (which could be exaggerated or fabricated), then the report about the Russian intelligence analysis (which is dependent on US intelligence collection of uncertain reliability), then the original context of the declassified documents (as hard evidence or simply examples of leads), and then the intention and appropriateness of the declassification itself. Unpacking each order of disinformation, their specific means of access and dissemination, and the associated possibilities for falsification and contestation reveals how broader narratives are dependent on the compilation of contested claim upon contested claim, stacking these claims into worldviews that have begun to rupture the US political system.

5. CONCLUSION

Round like a circle in a spiral, like a wheel within a wheel
Never ending or beginning on an ever-spinning reel

Alan and Marilyn Bergman, "The Windmills of Your Mind" (1968)

The relevance of second, third, and higher orders of disinformation will only increase as more experienced actors draw on the material, successes, and lessons of previous campaigns to construct new material. As Rid has demonstrated, Soviet active measures drew extensively on earlier controversies, even to the point of resurrecting previously debunked forgeries decades later.[40] We can expect this dynamic to play out on social media platforms and the internet, as what quickly become historic struggles over the factual record transform into the foundations of future narrative contestation. One of the most striking qualities of these ever-growing chains of disinformation is their reflexive nature – hence the resurrection of a popular 1960s song in the title of this paper and the quotation above. Indeed, the indirect consequences of these chains of higher-order disinformation, fracturing worldviews and exacerbating political polarisation, may themselves be a desired strategic effect of such operations.

More concretely, the specific policy implications of higher-order forms of disinformation can be divided into two kinds: defensive and offensive. Defensively, this approach reinforces scholarship indicating the limited utility of fact-checking services

---

[40] Rid, *Active Measures*.

in countering disinformation. While these services certainly have an important role to play in a turbulent communications ecosystem, many commentators have argued that they are unable to address broader narrative contestation – as in the example of the EEAS website considered earlier. We can now see why: to do so, fact-checking services would have to reverse-engineer multiple orders of disinformation, a time-consuming, resource-intensive process to say the least – and one likely to introduce its own biases. More problematically, a focus on higher orders of disinformation highlights that fact-checking services themselves are an attractive target for disinformation. In Tunisia, an Israeli PR company set up a fake fact-checking service before local elections, while in the UK the Conservative Party renamed its Twitter account "Factcheck UK" in the run-up to the 2019 election.[41] A growing wave of investigative journalism seeks to peel away such layers of misdirection – especially organisations such as Bellingcat, whose use of leaks has itself attracted some controversy – and so further analysis of the exploitation of fact-checking services would be a natural extension of the conceptual approach developed in this paper.

Offensively, the greater the salience of disinformation in international politics, the more all states – and other actors – will employ not just accusations of disinformation but also influence campaigns as a response to unwelcome international attention. Recent events illustrating this trend include China's response to the UK's withdrawal of the media license for a Chinese state-owned channel in February 2021. Chinese statements announced a reciprocal ban against BBC World News. The same statements denounced as "false information" the BBC's investigations of severe human rights violations against Uighurs in Xinjiang province (which in turn relied on leaked documents as well as interviews). This denunciation was backed up by a tightly coordinated influence campaign by government-linked accounts on Twitter.[42] In the same month, Saudi Arabia's furious response to the Biden administration's release of a report on the killing of Jamal Khashoggi not only branded the report itself as a disinformation operation but reinforced this message on Twitter using a network of bots like those Khashoggi worked against before his death.[43]

This response option is not limited to authoritarian states. Many militaries and intelligence agencies – including in the US and other NATO states – are openly considering more active responses to disinformation by adversaries along the lines of

---

41   Andy Carvin et al., "Operation Carthage: How a Tunisian Company Conducted Influence Operations in African Presidential Elections", *Atlantic Council*, June 5, 2020, https://perma.cc/AEY3-R3XU; Hannah Murphy and Alex Barker, "Conservative Party's 'FactcheckUK' Twitter Stunt Backfires", November 19, 2019, https://www.ft.com/content/0582a0d0-0b1f-11ea-b2d6-9bf4d1957a67.

42   Patrick Wintour, "China Bans BBC World News in Retaliation for UK Licence Blow", *Guardian*, February 11, 2021, http://www.theguardian.com/world/2021/feb/11/china-bans-bbc-world-news; Jacob Wallis and Albert Zhang, "Trigger Warning: The CCP's Coordinated Information Effort to Discredit the BBC" (Canberra: Australian Strategic Policy Institute, March 4, 2021), https://www.aspi.org.au/report/trigger-warning.

43   Craig Timberg and Sarah Dadouch, "When U.S. Blamed Saudi Crown Prince for Role in Khashoggi Killing, Fake Twitter Accounts Went to War", *Washington Post*, March 2, 2021, https://www.washingtonpost.com/technology/2021/03/02/saudi-khashoggi-twitter-mbs/.

those developed for other malicious actors in cybersecurity.[44] Such "counter-cyber" responses to disinformation include disrupting technical and social infrastructure, as reportedly occurred for the Internet Research Agency, a Russian "troll farm", before the US 2018 mid-term elections.[45] They also include clandestine social media campaigns, such as the one targeting Russia in the Sahel in December 2020, attributed by Facebook to the French military.[46] But this spectrum of responses also includes leaking adversaries' identities, tactics, and plans and (although this is not publicly stated) potentially including falsified or doctored information in these leaks. The utility of these operations must be evaluated carefully, not just in terms of the operations themselves as second- or third-order forms of disinformation, but also in terms of the potential for blowback – for the operations to be exposed by adversaries and incorporated into even higher order forms of disinformation.

In sum, this paper has argued that narrative contests involving repeated, escalating, and – crucially – *reflexive* accusations of disinformation on both sides are the norm rather the exception in international politics. The concept of higher orders of disinformation helps us to gain analytical purchase on such chains of successive and deeply disputed claims, and so – in a small way – contributes to the accurate diagnosis, and eventual amelioration, of a perennial problem.

44 Max Smeets, "U.S. Cyber Strategy of Persistent Engagement and Defend Forward: Implications for the Alliance and Intelligence Collection", *Intelligence and National Security* (February 15, 2020): 1–10, https://doi.org/10.1080/02684527.2020.1729316.

45 Catalin Cimpanu, "US Wiped Hard Drives at Russia's 'Troll Factory' in Last Year's Hack", *ZDNet*, February 28, 2019, https://perma.cc/763D-CEAY.

46 Nathaniel Gleicher and David Agranovich, "Removing Coordinated Inauthentic Behavior from France and Russia", *About Facebook*, December 15, 2020, https://about.fb.com/news/2020/12/removing-coordinated-inauthentic-behavior-france-russia/.

REFERENCES

Baezner, Marie. "The Use of Cybertools in an Internationalized Civil War Context: Cyber Activities in the Syrian Conflict". CSS Cyber Defense Project. Center for Security Studies, ETH Zurich, October 18, 2017.

Benkler, Yochai, Robert Faris, and Hal Roberts. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. New York: Oxford University Press, 2018.

Bok, Sissela. *Secrets: On the Ethics of Concealment and Revelation*. New York and Toronto: Vintage, 1989.

Branch, Jordan. "What's in a Name? Metaphors and Cybersecurity". *International Organization* 75, no. 1 (2021): 39–70. https://doi.org/10.1017/S002081832000051X.

Buchanan, Ben. *The Hacker and the State: Cyber Attacks and the New Normal of Geopolitics*. Cambridge, MA: Harvard University Press, 2020.

Campbell, David. *Writing Security: United States Foreign Policy and the Politics of Identity*. Manchester: Manchester University Press, 1998.

Carvin, Andy, Luiza Bandeira, Graham Brookie, Iain Robertson, Nika Aleksejeva, Alyssa Kann, Kanishk Karan et al. "Operation Carthage: How a Tunisian Company Conducted Influence Operations in African Presidential Elections". *Atlantic Council*, June 5, 2020. https://perma.cc/AEY3-R3XU.

Cimpanu, Catalin. "US Wiped Hard Drives at Russia's 'Troll Factory' in Last Year's Hack". *ZDNet*, February 28, 2019. https://perma.cc/763D-CEAY.

Cohen, Zachary, and Alex Marquardt. "Former CIA Director Accuses Intel Chief of Selectively Declassifying Documents to Help Trump". CNN, October 7, 2020. https://www.cnn.com/2020/10/06/politics/brennan-ratcliffe-declassifying-intelligence-clinton-russia/index.html.

Deibert, Ronald J. *Reset: Reclaiming the Internet for Civil Society*. Toronto: House of Anansi Press, 2020.

Desiderio, Andrew, and Daniel Lippman. "Intel Chief Releases Russian Disinfo on Hillary Clinton That Was Rejected by Bipartisan Senate Panel". *Politico*, September 29, 2020. https://www.politico.com/news/2020/09/29/john-ratcliffe-hillary-clinton-russia-423022.

Dunleavy, Jerry. "Obama Was Briefed on Unverified Russian Report Claiming Clinton Approved Plan to Tie Trump to Putin and DNC Hack". *Washington Examiner*, September 29, 2020. https://www.washingtonexaminer.com/news/obama-was-briefed-on-unverified-russian-report-claiming-clinton-approved-plan-to-tie-trump-to-putin-and-dnc-hack.

European Commission. "A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation". Luxembourg: European Commission Directorate-General for Communication Networks, Content and Technology, March 2018.

Facebook. "February 2020 Coordinated Inauthentic Behavior Report". *About Facebook*, March 2, 2020. https://about.fb.com/news/2020/03/february-cib-report/.

Fogarty, Stephen G., and Bryan N. Sparling. "Enabling the Army in an Era of Information Warfare". *Cyber Defense Review* 5, no. 2 (Summer 2020): 17–26.

Gleicher, Nathaniel, and David Agranovich. "Removing Coordinated Inauthentic Behavior from France and Russia". *About Facebook*, December 15, 2020. https://about.fb.com/news/2020/12/removing-coordinated-inauthentic-behavior-france-russia/.

Greer, Brian. "John Ratcliffe's Dangerous Declassification Game". *Lawfare*, October 7, 2020. https://www.lawfareblog.com/john-ratcliffes-dangerous-declassification-game.

Hulcoop, Adam, John Scott-Railton, Peter Tanchak, Matt Brooks, and Ronald J. Deibert. "Tainted Leaks: Disinformation and Phishing With a Russian Nexus". *Citizen Lab*, May 25, 2017.

Jamieson, Kathleen Hall. *Cyberwar: How Russian Hackers and Trolls Helped Elect a President – What We Don't, Can't, and Do Know*. New York: OUP USA, 2018.

Kerr, Jaclyn Alexandra, and Herbert Lin. "On Cyber-Enabled Information/Influence Warfare and Manipulation". *SSRN*, March 13, 2017.

Krebs, Ronald R. *Narrative and the Making of US National Security*. Cambridge University Press, 2015.

Lanoszka, Alexander. "Disinformation in International Politics". *European Journal of International Security* 4, no. 2 (2019): 227–248. https://doi.org/10.1017/eis.2019.6.

Merry, Sally Engle. "Rethinking Gossip and Scandal". In *Toward a General Theory of Social Control: Fundamentals*, edited by Donald Black, 271–302. Orlando and London: Academic Press, 1984.

Mueller, Robert S. *Report on the Investigation into Russian Interference in the 2016 Presidential Election*. Submitted Pursuant to 28 C.F.R. § 600.8(c). Washington, DC: US Department of Justice, March 2019.

Murphy, Hannah, and Alex Barker. "Conservative Party's 'FactcheckUK' Twitter Stunt Backfires", November 19, 2019. https://www.ft.com/content/0582a0d0-0b1f-11ea-b2d6-9bf4d1957a67.

Nimmo, Ben, Camille Francois, C. Shawn Eib, Lea Ronzaud, Rodrigo Ferreira, Chris Hernon, and Tim Kostelancik. "Secondary Infektion". *Graphika*, June 2020.

Pozen, David. "The Leaky Leviathan: Why the Government Condemns and Condones Unlawful Disclosures of Information". *Harvard Law Review* 127 (2013): 512–635.

Rid, Thomas. *Active Measures: The Secret History of Disinformation and Political Warfare*. New York: Profile Books, 2020.

Schneier, Bruce, and Henry Farrell. "Common-Knowledge Attacks on Democracy". Berkman Klein Center for Internet and Society, Harvard University, October 2018.

Sheth, Sonam. "Trump's Spy Chief Just Released 'Russian Disinformation' against Hillary Clinton that He Acknowledged May Be Fabricated". *Business Insider*, September 30, 2020. https://www.businessinsider.in/politics/world/news/trumps-spy-chief-just-released-russian-disinformation-against-hillary-clinton-that-he-acknowledged-may-be-fabricated/articleshow/78396299.cms.

Shires, James. "Hack-and-Leak Operations: Intrusion and Influence in the Gulf". *Journal of Cyber Policy* 4, no. 2 (2019): 235–256.

———. "The Cyber Operation against Qatar News Agency". In *The 2017 Gulf Crisis: An Interdisciplinary Approach*, edited by Mahjoob Zweiri, M. Mizanur Rahman, and A. Kamal. Berlin and Heidelberg: Springer Nature, 2020.

———. "The Simulation of Scandal: Hack-and-Leak Operations, the Gulf States, and U.S. Politics". *Texas National Security Review*, August 2020.

———. "Understanding the Tactics behind Hack-and-Leak Operations". *Atlantisch Perspectief* 4 (September 2020).

Singman, Brooke. "DNI Declassifies Brennan Notes, CIA Memo on Hillary Clinton 'Stirring up' Scandal between Trump, Russia". Fox News, October 6, 2020. https://www.foxnews.com/politics/dni-brennan-notes-cia-memo-clinton.

Smeets, Max. "U.S. Cyber Strategy of Persistent Engagement and Defend Forward: Implications for the Alliance and Intelligence Collection". *Intelligence and National Security* (February 15, 2020): 1–10. https://doi.org/10.1080/02684527.2020.1729316.

Timberg, Craig, and Sarah Dadouch. "When U.S. Blamed Saudi Crown Prince for Role in Khashoggi Killing, Fake Twitter Accounts Went to War". *Washington Post*, March 2, 2021. https://www.washingtonpost.com/technology/2021/03/02/saudi-khashoggi-twitter-mbs/.

US Department of Defense. "Summary: Department of Defense Cyber Strategy". Washington, DC, 2018.

Vilmer, Jean-Baptiste Jeangene. "The 'Macron Leaks' Operation: A Post-Mortem". Atlantic Council and IRSEM, June 2019.

Wallis, Jacob, and Albert Zhang. "Trigger Warning: The CCP's Coordinated Information Effort to Discredit the BBC". Canberra: Australian Strategic Policy Institute, March 4, 2021. https://www.aspi.org.au/report/trigger-warning.

Wintour, Patrick. "China Bans BBC World News in Retaliation for UK Licence Blow". *Guardian*, February 11, 2021. http://www.theguardian.com/world/2021/feb/11/china-bans-bbc-world-news.

Cyber Personhood

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

Keir Giles
Conflict Studies Research Centre
Northamptonshire, United Kingdom
keir.giles@conflictstudies.org.uk

Tassilo Singer
Consultant Manager
(Cyber Security & AI)
Atos Information Technology GmbH
Munich, Germany
tassilo.singer@atos.net

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

Abstract: In early 2020, the rapid adoption of remote working and communications tools by governments, companies, and individuals around the world increased dependency on cyber infrastructure for the normal functioning of States, businesses, and societies. For some, the urgent need to communicate whilst safeguarding human life took priority over ensuring that these communications tools were secure and resilient. But as these tools become firmly embedded in everyday life worldwide, the question arises whether they should be considered as critical infrastructure, or perhaps even something more important.

In a number of States, the critical importance of the environment for preservation of human life has been recognised by extending legal personhood – and thus, legal rights – to environmental entities. Countries such as Colombia, Ecuador, New Zealand, and India have granted legal rights to various rivers, lakes, parks, and nature in general. This paper explores the future possibility and cases where States may consider granting legal rights to other non-sentient but critically important entities. Looking into a future where human life becomes increasingly dependent upon highly interdependent systems in cyberspace, is there a possibility that these systems are granted personhood?

Remote work and its cybersecurity implications could lead to an entirely new recognition of the importance of cyberspace dependencies and, consequently, a new

legal treatment. Against the backdrop of extended debate on the legal regulation of cyberspace, including the law of armed conflict, this would raise even more complex legal considerations, especially in the light of cross-border dependencies and systems that affect multiple jurisdictions.

By way of cyber biomimicry, this paper adopts a blue-sky conceptual approach to studying policy considerations and potential implications if highly interdependent cyber systems in the distant future are granted the same protections as elements of the environment.

1. INTRODUCTION

Under Canadian and U.S. environmental law, rivers, parks, and other natural resources upon which life depends do not have standing in their respective jurisdictional courts. Instead, in order for there to be standing, harm to any of these natural features must have resulted in injury to human beings. But what if natural resources were widely recognised in courts and had legal rights, with injuries to these natural resources recognised as crimes with victims in and of themselves?

If so, could this be extrapolated to a distant future where highly interdependent resources in cyberspace upon which life depends are also recognised, on the basis that these too are dynamic systems that have standing so that courts can recognise their injuries? This concept may appear unlikely, but so did the idea of environmental personhood decades ago, and today it is reality. In a world where our dependence on cyber systems is ever increasing, the idea of States granting cyber personhood to highly interdependent cyber systems of the future could be a logical progression of a number of current trends.

This paper examines environmental personhood and how a small number of States have granted it to certain natural resources. Through examples, we then describe the term "cyber personhood," align it to the precedent set by environmental personhood, present candidates for cyber personhood, and identify where we believe cyber personhood could not apply and where it may.

Finally, we examine certain policy considerations and potential implications of cyber personhood and provide our thoughts on the wider adoption of this concept.

In order to digest the content presented in this paper, we urge the reader to (1) look far into the future to help visualise these highly interdependent cyber systems and (2) not consider current cyber systems as candidates for cyber personhood. To help standardise our discussion across the political, policy, legal, and technological domains, we present the following definitions.

- Cyberspace: "The environment formed by physical and non-physical components to store, modify, and exchange data using computer networks."[1]
- Cyber infrastructure: "The communications, storage, and computing devices upon which information systems are built and operate."[2]
- Critical infrastructure:
  i. "Physical or virtual systems and assets of a State that are so vital that their incapacitation or destruction may debilitate a State's security, economy, public health or safety, or the environment."[3]
  ii. "…infrastructure sectors whose assets, systems, and networks, whether physical or virtual, are considered so vital to the United States that their incapacitation or destruction would have a debilitating effect on security, national economic security, national public health or safety, or any combination thereof."[4]
- Cyber system: "One or more interconnected computers with associated software and peripheral devices. It can include sensors and/or (programmable logic) controllers, connected over a computer network. Computer systems can be general purpose (e.g. a laptop) or specialised (e.g. the 'blue force tracking system')."[5]
- Highly interdependent cyber systems of the future: Defined by examples in the following section.

2. HIGHLY INTERDEPENDENT CYBER SYSTEMS

We are already on the brink of a future in which we depend so much on key cyber systems that governments, societies, corporations, and individuals are, in some cases, unable to function without them. Current trends indicate that this dependence on always-on, always-reliable cyber systems will deepen. During the coronavirus pandemic, without the ability to operate remotely, many more companies would have failed and more individuals relying on their services would have suffered. In the spring of 2020, governments and companies scrambled to increase secure remote

---

1   Michael N. Schmitt (ed.), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*, (Cambridge: Cambridge University Press, 2017), 564.
2   Ibid.
3   Ibid.
4   "Critical Infrastructure Sectors," Cybersecurity Infrastructure Security Agency (CISA), https://www.cisa. gov/critical-infrastructure-sectors [accessed 8 March 2021].
5   Schmitt, *Tallinn Manual 2.0*, 564.

access capacities and adopt remote voice and video communication methods. For some, these voice and video communications systems now rely on a complicated mix of on-premise systems, service providers, cloud hosting providers, and Internet access to residences. This delicate balance of a service offering relies on the availability of each service component within and is an example of our existing dependency on always-on and always-reliable cyber systems. So far, the roles and responsibilities remain clear.

Near-future examples demonstrating this include fully autonomous vehicles (without steering wheels) and systems that are critically dependent on synchronised time signals, not only for waking you up in the morning and scheduling your day but also for key tasks such as ensuring your digital identity and encryption for connecting to your common services in cyberspace. The roles and responsibilities in such a technological system begin to blur, as they all depend on *time*. As an example, given that satellite time signals can be manipulated or jammed,[6] common services that depend on time, such as locations on maps and certificate expirations that influence identities and cryptography, could cease to function as intended. These temporary effects, described in the example, demonstrate the potential for harm to the operations of systems dependent on a synchronised time signal.

Further into the future, societies may rely on cyber systems based on emergent phenomena in complexity theory systems,[7] or cyber physical systems[8] managed entirely by artificial intelligence (AI) systems, where the original human-written algorithms of the system are regularly rewritten by the learning process of the system itself. The closer cyber systems get to sentience, the more rational it becomes to treat them as legal entities in their own right, capable not only of suffering harm but also of taking decisions that cause harm independently of human input.

Now consider multiples of these future cyber systems being highly interdependent on each other, where they feed and receive data from each other and also consume each other's deeply nested computing capacities. These systems would be managed by companies or governments and potentially poorly designed by individuals, like many

---

6   Peter Danilov, "GPS Jamming Still Causing Problems in Finnmark," *High North News*, 19 November 2020, https://www.highnorthnews.com/en/gps-jamming-still-causing-problems-finnmark [accessed 8 March 2021].

7   Paul Cilliers, *Critical Complexity: Collected Essays*, ed. Rika Preiser (Berlin: Walter de Gruyter GmbH, 2016). We understand Paul Cilliers' view of a complex system to be a large number of elements (which can be simple), interacting dynamically and nonlinearly using feedback loops, where the system behaviour is determined by these interactions. Such systems are adaptive, reorganising their internal structure without intervention by outside agents.

8   Claire Vishik, Mihoko Matsubara, and Audrey Plonk, "Key Concepts in Cyber Security: Towards a Common Policy and Technology Context for Cyber Security Norms," in *International Cyber Norms: Legal, Policy and Industry Perspectives*, ed. Anna-Maria Osula and Henry Rõigas (Tallinn: NATO CCD COE Publications, 2016), 228–229. Cyber physical systems are smart systems that include co-engineered interacting networks of physical and computational components. Thanks to their highly interconnected nature, they are an excellent example of complex systems in which the behavioural sum is far more than its parts.

other systems today (e.g. a great many commercial software packages). The roles and responsibilities that were once complicated will become complex. If, in the future, these highly interdependent cyber systems were to become temporarily unavailable or significantly harmed, it could impact a State's (or various States') ability to deliver healthcare or to maintain international obligations, cause a shutdown of the economy, and possibly cause civil unrest. If, as you read this, you find yourself trying to align our description of these highly interdependent cyber systems with Critical National Infrastructure or trying to align to software and/or services you may use today, then we ask that you look much further into the future and set aside any alignment to services that currently exist.

In the following section, we explore the position States have taken with respect to environmental personhood and later align this behaviour and thought to our example-based definition of highly interdependent cyber systems of the future to discuss cyber personhood.

3. ENVIRONMENTAL PERSONHOOD

It is simple to understand a corporation having its own rights as a legal entity, and thereby corporate personhood. These corporate entities can enter into contracts, own properties, and be recognised as legal persons in courts. But in addition to corporations, natural resources in certain States have been granted personhood rights.

The germination of environmental personhood is credited to the 1972 paper "Should Trees Have Standing? Toward Legal Rights for Natural Objects" by Christopher D. Stone.[9] The paper proposed giving legal rights to rivers, oceans, forests, or any natural environmental systems.[10] He referred back to a time when discussing rights for corporations, women, and others had seemed unthinkable.[11] He went on to describe how corporations do not have rights similar to those of a legal person and how certain persons such as inmates or children have limited rights.[12] He argued that "holders of legal rights" must satisfy all of the following three criteria:[13]

1. "[They may] institute legal actions at [their] behest;"
2. "In determining the granting of legal relief, the court must take injury to [them] into account; and"
3. "Relief must run to [their] benefit."

9   Christopher D. Stone, "Should Trees Have Standing? Toward Legal Rights for Natural Objects," *Southern California Law Review* 45 (1972): 450–501.
10   Ibid., 456.
11   Ibid., 451.
12   Ibid., 455.
13   Ibid., 458.

Stone suggested that these natural environmental systems be assigned legal guardians who could advocate for the rights of these systems.[14] In his blue-sky paper, he suggested not only rights but also liabilities, using the example of a trust fund which compensates those who suffered damages from floods.[15]

Since Stone's paper, some nations have shifted slightly from anthropocentric views toward biocentric ones, adopting environmental personhood in a number of different ways. Several papers have been published regarding the interpretation and challenges of State laws regarding environmental personhood. Examples include:

- **Bangladesh:** In 2019, Bangladesh granted legal personhood rights to all of its rivers, with legal guardianship assigned to the National River Conservation Commission.[16]

- **Bolivia:** In 2010, Bolivia passed a "Law on the Rights of Mother Earth" (Ley de Derechos de la Madre Tierra), thereby granting her, a living system, legal personhood rights.[17]

- **Colombia:** In 2016, Colombia's Constitutional Court granted legal personhood rights to the Atrato River (Rio Atrato) basin under joint guardianship of the government and the indigenous community living in the basin.[18] In 2018, Colombia's Supreme Court recognised the rights to the Amazon River and its surrounding ecosystem, reaching a unique decision involving multiple stakeholders to safeguard the life and health of Colombia's Amazon (Amazonas Colombiano).[19]

- **Ecuador:** Leading the charge in 2008, Ecuador's constitution recognises legal personhood for "Mother Nature" (Panchamama) with rights "to exist, persist, maintain and regenerate its vital cycles, structure, functions and its processes in evolution." Any person or persons can petition on her behalf.[20]

---

14 Ibid., 464.
15 Ibid., 481.
16 Supreme Court of Bangladesh, Human Rights and Peace for Bangladesh v. Bangladesh and Others (HRPB v. Bangladesh), Writ Petition 13989/2016 of 7 November 2016.
17 "Bolivia Law of the Rights of Mother Earth," Law 071 (2010).
18 Justice Studies et al. v. Presidency of the Republic et al., Constitutional Court of Colombia, Judgment T-622/16, https://www.corteconstitucional.gov.co/relatoria/2016/t-622-16.htm [accessed 8 March 2021].
19 Supreme Court of Colombia, Judgement STC 4360-2018 of 5 April 2018, https://cortesuprema.gov.co/corte/wp-content/uploads/2018/04/STC4360-2018-2018-00319-011.pdf [accessed 8 March 2021].
20 Constitución Política de la República del Ecuador, Constitución 2018, Art. 71–74.

- **India:** In 2017, India's Uttarakhand High Court granted legal personhood rights to two rivers, the Ganges and the Yamuna, their respective Gangotri and Yamunotri glaciers, and other natural objects[21] in the State of Uttarakhand under the guardianship of Uttarakhand, the State in which the rivers originate. Later that same year, India's Supreme Court issued a stay of the Uttarakhand High Court's 2017 decision.[22] In March 2020, the Punjab and Haryana High Court granted Sukhna Lake personhood rights.[23]

- **New Zealand:** New Zealand granted legal personhood rights to the Te Urewera National Park in 2014,[24] the Whanganui River in 2017,[25] and Mount Taranaki in 2017,[26] with legal guardianship assigned to the Crown, the Whanganui people, and eight Māori tribes, respectively.

We recognise that with the exception of India, these States may not be globally perceived as legal opinion defining States. With the further exception of Ecuador and Bolivia, we also recognise that not all aspects of the State's environment are granted legal personhood and that only specific rivers, forests, and parks have been granted legal personhood. It is most likely for these reasons that environmental personhood is not a rule in public international law or included in customary international law.

Rather than explore the legal constructs developed to create a concept of environmental personhood, this paper builds on the established notion to consider a distant future where some States grant personhood rights to highly interdependent cyber systems. It is with this frame of reference that we propose cyber personhood.

---

[21] High Court of Uttarakhand, Mohammad Salim vs. State of Uttarakhand & others, Writ Petition (PIL) No. 126 of 2014, 20 March 2017: "§19. Accordingly, while exercising the parens patriæ jurisdiction, the Rivers Ganga and Yamuna, all their tributaries, streams, every natural water flowing with flow continuously or intermittently of these rivers, are declared as juristic / legal persons / living entities having the status of a legal person with all corresponding rights, duties and liabilities of a living person in order to preserve and conserve river Ganga and Yamuna."
High Court of Uttarakhand, Lalit Miglani vs. State of Uttarakhand & others, Writ Petition (PIL) No. 140 of 2015, 30 March 2017: "We, by invoking our parens patriæ jurisdiction, declare the Glaciers including Gangotri & Yamunotri, rivers, streams, rivulets, lakes, air, meadows, dales, jungles, forests wetlands, grasslands, springs and waterfalls, legal entity / legal person / juristic person / juridicial person / moral person / artificial person having the status of a legal person, with all corresponding rights, duties and liabilities of a living person, in order to preserve and conserve them. They are also accorded the rights akin to fundamental rights / legal rights."

[22] Supreme Court of India, State of Uttarakhand and Others v. Mohammed Salim and Others, Special Leave to Appeal (C) No. 016879/2017, Order dated 7 July 2017.

[23] Punjab and Haryana High Court, Court on its own motion v. Chandigarh Administration, CWP No. 18253 of 2009 and other connected petitions of 2 March 2020.

[24] Parliament of New Zealand, "Te Urewera Act 2014," Royal Assent 27 July 2014.

[25] Parliament of New Zealand, "Te Awa Tupua (Whanganui River Claims Settlement) Act 2017," Royal Assent 20 March 2017.

[26] "Taranaki Maunga," signed 20 December 2017, https://www.govt.nz/browse/history-culture-and-heritage/treaty-settlements/find-a-treaty-settlement/taranaki-maunga [accessed 8 March 2021].

4. CYBER PERSONHOOD

Our paper suggests that in the distant future, States may grant or consider granting certain highly interdependent cyber systems legal personhood rights, in a manner similar to how States have granted certain natural environment systems environmental personhood. In common with environmental personhood, the rights and liabilities of these cyber systems would, and most likely should, vary from system to system.

We propose the following definition of cyber personhood: the granting of legal-person rights to a highly interdependent cyber system under legal frameworks whereby the highly interdependent cyber system would have legal standing to claim injuries and remain accountable for any injuries it may cause.

The notion of granting legal personhood to a computer-based system may seem radical and exotic at present, but far less so than the idea of environmental personhood did in 1972. While environmental concerns have slowly achieved broad acceptance despite being stigmatised by industry-minded or politically motivated interests, dependency on cyber systems is developing much more rapidly. Where it took over 35 years for environmental personhood to take hold in Ecuador, it is possible that cyber personhood will mirror the velocity of acceptance of cyber systems, greatly reducing the time required to arrive at appropriate legal changes to recognise the new reality, or dismiss it.

A. Candidates for Cyber Personhood

Just as with corporate legal entities or inmates, the highly interdependent systems of the future would probably fall into a category of their own, requiring different treatment, including in terms of their rights and obligations, as well as forms of ownership and oversight. To help understand the types of systems that may be considered for cyber personhood, we have categorised them as follows. For each scenario, our focus remains on the highly interdependent cyber systems of the future, States' and their societies' inability to function without them, and existing legal constructs that may apply, which essentially negates the concept of cyber personhood for the first two candidates described below.

1. **Individually owned cyber systems (personal):** Many people have small networks in their homes providing connections between devices within their home, such as their computers, mobile devices, and televisions, and fringe devices such as refrigerators, toasters, coffee makers, door locks, and other Internet of Things objects. This candidate is not a highly interdependent cyber system but can be impacted if it is reliant upon upstream highly interdependent cyber systems that are no longer available. Nevertheless, if

this example's services were to be unavailable, it would not gravely impact State ability to function, and any damages can be claimed by the owner. *We believe such personal systems are not candidates for cyber personhood*.

2. **Corporate- or State-owned cyber systems (single entity):** These systems are required by corporate or State entities to operate their daily business, and if they were made unusable, the impact would be localised to their operations. *These systems would likely not be granted cyber personhood*, as any damages to them can be claimed by the owner, and any damages from the system can be paid by the owner.

3. **Multi-entity-owned cyber systems in a single jurisdiction:** In this instance, several national companies combined with or without the State's owned systems leverage their respective cyber services to jointly offer services from highly interdependent cyber systems to residents of a single jurisdiction. An example of this would be a nation that is able to provide cyber services to its residents thanks to its extensive sovereign cyber capabilities at State and/or corporate levels. These interdependent cyber systems could maintain separate accountabilities, leaving each entity responsible for their portion of the system. It may also simply fall under the responsibility of the State, especially when trying to limit control from larger and more powerful corporations such as Alphabet, Facebook, Amazon, and Microsoft. Alternatively, States may implement a private-public partnered governing body to govern the system as a single unit, especially when the boundaries of the individual units within the system become difficult to ascertain. For example, what would happen if one entity or service provider within the overall highly interdependent cyber system decides to stop providing its service, thereby adversely impacting all entities and the overall service to the residents of the State? *We believe it is possible for States to grant such a system cyber personhood*.

4. **Multi-entity-owned cyber system across multiple jurisdictions:** Building upon the previous candidate, consider several multinational companies and/ or several States that jointly offer a service through a highly interdependent set of cyber services to the residents of several jurisdictions, including jurisdictions beyond their own with complex and deeply nested roles and responsibilities. Depending on the public's reliance on the services of the system and the level of impact to the public when the services offered through the system are lost, *we believe such systems may be considered by some States to be deserving of cyber personhood*.

5. USE AND POLICY CONSIDERATIONS

In addition to candidates of cyber personhood that would require new legal treatment as described above, specific instances of actions affecting (or indeed carried out by) highly interdependent cyber systems would require careful consideration when establishing a conceptual framework for cyber personhood. Had we had the foresight to strategise or "pre-think" our handling of the coronavirus pandemic, globally we would have been in a better position than where we arrived a year later. This paper suggests that States pre-think the idea of cyber personhood to help them decide how they would respond if certain States adopt such a position.

The following is a non-exhaustive list of considerations influencing rights and obligations of legal persons that would have a distinctive impact when applied to highly interdependent cyber systems of the future that States and their societies would be unable to function without.

- **Injuries:** the nature of highly interdependent cyber systems of the future, existing in the physical world yet managing data in the virtual one, means that the potential for damage caused by cyber systems also extends across multiple domains. In the data sphere, highly interdependent cyber systems could be liable and receive relief for breach of confidentiality, damage to integrity, or breach of access, or damage or destruction of systems or data. In the physical world, harm could be caused to any system – including life support systems – which is dependent on the network for its correct functioning. Interdependencies introduce further complexity when, for instance, one entity's components of the highly interdependent cyber system cause harm to another entity's components of the same system, where one or both has been designated as a legal person.

- **Cyber attack (outside of armed conflict):** In the future, when a highly interdependent cyber system becomes the victim of a cyber attack, its rights and duties depend on the existence of an organisational body and the prevailing degree of organisation, as well as on its "legal" recognition by States on a national and international level.

  - If the highly interdependent cyber system has been granted cyber personhood by a State in which its rights can be invoked, then those rights (and duties – like a duty to notify/report authorities on serious breaches) can be exercised in front of a national jurisdiction. Furthermore, the executive branch could be asked for assistance in the form of, for example, preventive protection or services such as

attribution sourcing and security monitoring. As a result of such a legal remedy, the most basic expectation would be a return to the pre-attack status of the highly interdependent cyber system.

○ On the international level, a cyber attack could result in a demand by interested parties[27] to protect the system, restore it to its pre-attack status, or to retaliate with sanctions. Additionally, if the rules of the customary law on State responsibility for States can be transferred to a highly interdependent cyber system, "third States" with common interests would be permitted to invoke them and could offer assistance.[28] Due to the necessity and/or criticality of the highly interdependent cyber system we have proposed, it is suggested that States also consider transferring these rules to such a non-State entity in order to support international laws of State responsibility.

- **Cyber attack (armed conflict):**[29] The protection under the law of armed conflict depends on how the highly interdependent cyber system is qualified. If it is equivalent to critical infrastructure, it enjoys a high standard of protection.[30] If the cyber system is used exclusively for civilian purposes, it is qualified as a civilian object and thus also protected.[31] Unfortunately, if a highly interdependent cyber system is abused by one party to an armed conflict, it could lose its status. When it becomes a civilian object used for military purposes, it can be qualified as a military objective.[32] Since an attack on a military objective results in a military advantage, a cyber attack on such a highly interdependent cyber system would be lawful. Legal reasons justifying protection could be an agreement of States on the neutrality of highly interdependent cyber systems or to qualify them as a "digital" non-defended locality.[33] Even more interestingly, due to the similar understanding

---

[27] Examples of interested parties include, but are not limited to, (1) the recognising States, (2) the users, (3) international organisations, (4) the systems' organisational body, and the like.

[28] UN ILC, "Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries" (2001), GAOR 56th Session Supp 10, 43; Art. 48: "1. Any State other than an injured State is entitled to invoke the responsibility of another State in accordance with paragraph 2 if: (a) the obligation breached is owed to a group of States including that State, and is established for the protection of a collective interest of the group; or (b) the obligation breached is owed to the international community as a whole." The cyber system could reflect the "collective interest" and/or the protection of the cyber system is "owed to the international community…" due to its criticality.

[29] Schmitt, *Tallinn Manual 2.0*, 415. Rule 92 of *Tallinn Manual 2.0* defines a cyber attack as "a cyber operation, whether offensive or defensive, that is reasonably expected to cause injury or death to persons or damage or destruction to objects" (415).

[30] Protections similar to those under Art. 54 and 56 Additional Protocol I to the Geneva Conventions, 1977.

[31] "Civilian objects shall not be the object of attack or of reprisals. Civilian objects are all objects which are not military objectives": Art. 52 (1) Additional Protocol I to the Geneva Conventions, 1977.

[32] As per Art. 52 (2) and within the limits of Art. 52 (3) Additional Protocol I to the Geneva Conventions, 1977; Yoram Dinstein, *The Conduct of Hostilities under the Law of International Armed Conflict* (Cambridge: Cambridge University Press, 2016): 104, 111–114.

[33] As per Art. 59 Additional Protocol I to the Geneva Conventions, 1977.

of the wording, they could be qualified as a demilitarised zone by agreement of the State parties to the conflict.[34]

- **Obligations:** The delineation of responsibility between the creator or designer of a system and the system itself will need to be strictly determined. The system will need to demonstrate that it has a certain degree of autonomy in order for responsibility for its actions to not wholly be that of its designers, programmers, or (in the case of AI) trainers. Under a governing body, those maintaining the system will also have an ongoing degree of responsibility for any changes introduced in its functioning.

- **Liabilities:** Where injuries have been caused by a system, the question arises of how these are to be recompensed, and whence funding is to be derived in order to compensate the victim. Equivalent to legal persons and corporations, highly interdependent cyber systems of the future will have the option to purchase liability insurance (or remain self-insured) and use trust funds to support injury costs against them. Users of the highly interdependent cyber systems could be charged a fee for services and/or the highly interdependent cyber systems could receive their funding from State-collected taxes, which would fund the maintenance and operations of the services along with costs for insurance policies and trust funds.

- **Representation and/or guardianship:** Until such time as systems can argue their own cases in courts of law, they must necessarily be represented by advocates in the same way as human or corporate plaintiffs or defendants. However, it also follows logically from personhood that sentient systems may also seek representation in corporate and political as well as legal systems in the same manner as any disenfranchised group of humans has sought to organise order to ensure their own rights – whether through a guild, or trades union, or by seeking political influence at a local or State level.

When selecting and assigning guardianship, States will likely consider the challenges of industry-driven or politically motivated interests. The system may be assigned multiple parties to act as a committee with guardianship responsibilities of the system, possibly consisting of preservation or advocacy groups, involved corporations, and a political seat (e.g. minister of the highly interdependent cyber system), similar to what has been assigned for natural environment entities. Certain international organisations could serve as a possible model.

---

34 As per Art. 60 Additional Protocol I to the Geneva Conventions, 1977.

○ International organisations as a comparable concept: A related practice-oriented solution for multi-jurisdictional systems can be an international governing body and/or international organisation (IO) deriving from and in accordance with international law. The required pressure and/or need to organise certain IT issues on an international level is comparable and similar to the ICANN (Internet Corporation for Assigned Names and Numbers, whose duty is to maintain important databases related to namespaces and numerical spaces of the Internet) or the ISO (International Organisation for Standardization, an association under Swiss law), which, however, are not governed precisely like an IO in the international law sense. It is therefore suggested that the practical idea of ICANN et al. be merged with the concept of an IO in international law. This might be explicitly suitable for a sophisticated AI complex.

To establish an international organisation, an agreement by at least two States in the form of an international treaty is required. In this treaty, the subject matter will be defined as well as its and the participating States' rights, duties, and funding.[35] From a practical perspective, it would be necessary to define the area of applicability precisely and thereby to determine and differentiate the highly interdependent cyber systems which are governed, guarded, and represented by the IO.

The creation of an international entity for a particular highly interdependent cyber system would entail the need for its own governance mechanisms. Furthermore, the integration of such a legal personhood in practice (i.e. procedural and representative questions) could be challenging; it could be addressed in a similar manner to existing specific IOs. On the other hand, the IO solution offers a clear and transparent framework based on States' consensus to govern a grey area and to answer legal and practical needs. Finally, particular advantages gained by creating this international entity could be:

- The monitoring and observance of (digital) human rights (e.g. with a view to surveillance or big data AI);
- A fair and equal share of high-level technology (e.g. for developing States);
- To keep critical communication and information infra-structure worldwide functioning (as a backbone);

---

35   Reparation for injuries suffered in the service of the United Nations (Advisory Opinion), (1949) ICJ Rep 174.

- ▪ Shared responsibility and shared burdens with a view to sustainability (to prevent environmental damage, or to foster decarbonisation); and/or
- ▪ A common control and reciprocal acceptance of a pivotal technology (sophisticated, eventually somewhat dangerous AI).

6.   CONCLUSION

The information revolution has already brought about profound changes in the lives of most humans and in what is considered normal and natural human behaviour. The pace of this change continues to increase, and to a greater extent than in previous periods of human history, legal practice is considered only after the systems are already in place. The extent to which the development of cyber systems and capabilities has outpaced legal norms is demonstrated not only by the constant need to update domestic computer and information legislation[36] to reflect new uses and capabilities for information and communications technologies but also by the ongoing discussions of the nature of cyber activities and what constitutes a "cyber attack" between States.[37]

Our paper is written to help States pre-think the concept of cyber personhood. The example of environmental personhood cited above provides a template for consideration of whether cyber personhood is a viable means for ensuring that the legal treatment of highly interdependent cyber systems of the future remains both relevant and fit for purpose, and sufficiently flexible to accommodate as-yet-unforeseen developments in the relationships between humans and computing devices. Christopher Stone's 1972 paper first proposing environmental personhood came at a very early stage in the development of mass awareness of the vulnerability of the environment, and of its need for protection, based, not least of all, on its critical importance for sustaining human life. The process of achieving widespread acceptance of the notion that corporate profit and individual convenience needs to be balanced against environmental protection was a long one, and in some areas is still not complete. However, we believe that events such as the coronavirus pandemic will accelerate the analogous process for cyber systems by emphasising the essential and irreplaceable nature of highly interdependent cyber services for the functioning of future societies.

---

36   Alex Scroxton, "Security Pros Fear Prosecution under Outdated UK Laws," *Computer Weekly*, 20 November 2020, https://www.computerweekly.com/news/252492416/Security-pros-fear-prosecution-under-outdated-UK-laws [accessed 8 March 2021].

37   Sydney J. Freedberg Jr. and Theresa Hitchens, "Calling SolarWinds Hack 'Act of War' Just Makes It Worse," *Breaking Defense*, 21 December 2020, https://breakingdefense.com/2020/12/calling-solarwinds-hack-act-of-war-just-makes-it-worse [accessed 8 March 2021].

The legal regime governing actions against, through, or by computer networks will inevitably develop and change, evolving significantly from its current state. It may be that cyber personhood is not the concept through which legal mechanisms accommodate the new reality of critical human dependence on online services. But the example of environmental legislation argues strongly that this path could be considered a key means of resolving substantial challenges to applying existing legal regimes to cyber rights and responsibilities by way of cyber biomimicry.

Studying and remaining aware of potential future scenarios enables us to better position ourselves to withstand them. For many reasons, environmental personhood is not widely accepted or recognised; however, it may be that cyber personhood is embraced as highly interdependent cyber systems become indispensable to governments, societies, corporations, and individuals.

The concept of cyber personhood is not so far removed from possibility and deserves discussion, particularly as the tools to govern it are already available. The questions that remain are: which cyber systems will develop the criticality and complexity deemed to be worthy of being governed under international law, and which countries are bold enough to make this concept a reality?

Explainable AI for Classifying Devices on the Internet

Artūrs Lavrenovs
NATO CCDCOE
Tallinn, Estonia
arturs.lavrenovs@ccdcoe.org

Roman Graf
Accenture GmbH
Vienna, Austria
roman.graf@accenture.com

Abstract: Devices reachable on the Internet pose varying levels of risk to their owners and the wider public, depending on their role and functionality, which can be considered their class. Discussing the security implications of these devices without knowing their classes is impractical. There are multiple AI methods to solve the challenge of classifying devices. Since the number of significant features in device HTTP response was determined to be low in the existing word-embedding neural network, we elected to employ an alternative method of Naive Bayes classification. The Naive Bayes method demonstrated high accuracy, but we recognise the need to explain classification results to improve classification accuracy.

The black-box implementation of Artificial Neural Networks has been a serious concern when evaluating the classification results produced in most fields. While devices on the Internet have historically been classified manually or using trivial fingerprinting to match major vendors, these are not feasible anymore because of an ever-increasing variety of devices on the Internet. In the last few years, device classification using Neural Networks has emerged as a new research direction. These research results often claim high accuracy through the validation employed, but through random sampling there always occur devices that cannot be easily classified, that an expert intuitively would classify differently. Addressing this issue is critical for establishing trust in classification results and can be achieved by employing explainable AI.

To better understand the models for classifying devices reachable on the Internet and to improve classification accuracy, we developed a novel explainable AI method, which returns the features that are most significant for classification decisions. We employed a Local Interpretable Model-Agnostic Explanations (LIME) framework to

explain Naive Bayes model classification results, and using this method were able to further improve accuracy with a better understanding of the results.

1. INTRODUCTION

With billions of devices connected to the Internet, it is not a question of whether the devices will be compromised or abused but when. Hundreds of millions of devices that are publicly reachable on the Internet are particularly vulnerable. Unsophisticated attackers can freely communicate with these devices and exploit configuration weaknesses or unpatched publicly disclosed vulnerabilities. Even if there are no currently disclosed vulnerabilities, these can appear at any time in the future. The classification of devices on the Internet has emerged in the last few years as an important research topic in the context of cyber security. Researchers and defenders have to understand new threats quickly and precisely in order to respond swiftly. Longstanding issues have to be understood so as to identify and address the root causes.

What class of devices has been compromised to create the latest Internet of things (IoT) botnet? What classes of devices have been abused for decades for distributed denial-of-service (DDoS) attacks? What devices receive a few anomalous network traffic flows from our network? These are just a few of many questions researchers and cyber security professionals have to answer. A pattern in the questions can already be observed inquiring about either large sets of devices or a few individual ones. Traditionally, this has been addressed either by applying a limited set of static classification rules or manual investigation by an expert. The increase in the number of devices is accompanied also by an increase in their heterogeneity. Static rules cannot keep up with this trend; therefore, the precision and also suitability of this method is decreasing. Expert availability is limited, and time is valuable – the automation of expert knowledge is the holy grail of AI application in the cyber security domain.

Expert knowledge, especially in cyber security, stretches far beyond applying standard tools and techniques. An expert's intuition is built upon years of experience, and the ability to validate predictions. An expert understands that a cloud computing network should not have many ICS devices present on it. And the few that might have a purpose would be specific to the infrastructure of the data centre. In comparison, even a sophisticated ML classifier classifies a large number of IoT devices in commercial

hosting and cloud networks [1]. Is this only an issue of lacking a network name or type (e.g. residential, commercial, cloud) as a feature? This feature can easily be added [2], but it is still not a guaranteed fix. While this is an obvious anomaly that could be easily identified and addressed by an expert, it is unclear how prevalent the misclassification is in the current research body, which stems from the black-box approach of the ML classification.

Feature selection is generally based on expert knowledge. Experts attempt to transfer their knowledge and characterise their intuition in the specific domain similar to how static rules would be created. Features for device classification with HTTP interfaces include directly identifiable keywords (the typical way to define a static rule), the behaviour of the responses, exclusion conditions, and the statistical properties of the responses [3], [4]. Expert feature decisions can be based on external sources of information; for example, a detailed scan of the device, which makes validating classification results much harder. In all the cases, the validity of feature selection can be questioned. While it is unfeasible for the expert to define all the less common features, experts can easily miss relationships between common features or put too much emphasis on some. Without explainable classification, feature selection and tuning can become overly reliant on the initial expert input.

The trustworthiness of classification even before adversary attempts are considered is the main obstacle to adoption in production. While the continuous improvement of the classification of identified issues and increasing precision is the expected progression in research, a single misclassification can be catastrophic in a cyber defence setting. Another advantage of the suitability of this research is that the level of precision is approaching expert knowledge for large sets, which has never before been possible.

The contributions of this paper include applying explainable AI to the problem of the classification of devices for the first time in published literature and bridging the gap between expert knowledge and automated classification.

Section 2 reviews related work, and Section 3 describes the application of explainable AI for the problem of device classification. Section 4 analyses the classification explanation for a random device from each defined class. Section 5 provides an overview of the overall classification results, and Section 6 provides final conclusions.

2. RELATED WORK

Scanning the Internet for specific devices or protocols is an established practice in security research. This type of research in itself has no novelty in respect to the classification process. Assumptions can be made that a device with a known open port corresponding to a non-generic protocol is serving a role that could be easily classified. Further validation by executing protocol communications can be conducted and data potentially useful for classification extracted. This methodology is effective for locating high-impact devices that are running specific protocols (commonly ICS) for the purpose of disabling public access. Mirian *et al.* scanned the Internet for common industrial protocols while identifying the discrepancy between open ports and the ability to handle respective protocol handshakes [5]. Dahlmanns *et al.* explore the security issues for the publicly reachable industrial protocol OPC UA [6]. Feng *et al.* automated IoT classification rule generation [7].

The privileged observer can identify traffic passing through network routers. The basic properties of port and protocol communication can be similar, while active communication requires sophisticated fingerprinting. This approach might make it possible to identify devices that are not publicly reachable but are actively communicating, while at the same time, it might miss devices that are not actively sending packets. Nawrocki *et al.* utilised IXP and ISP vantage points to identify common industrial protocols while still being challenged by traffic classification [8].

The research into AI classification consists of the same two vantage point approaches. The main challenge is identifying features and labelling sufficient training sets. Yang *et al.* identified and classified ICS and IoT devices extracting features and fingerprints from multiple communication layers [9]. Augmenting this with automated rule generation saved a significant amount of work for labelling the training set. Lavrenovs *et al.* trained classifier targeting interfaces based on generic HTTP protocols [2]. Privileged network observer classifiers are commonly trained on labelled data either from a laboratory network [10] or a campus network [11], [12]. Yadav *et al.* provide a systematic categorisation of ML augmented techniques for fingerprinting IoT devices [13].

Due to the fact that many AI models follow the black-box approach in terms of result transparency, research in the explainable AI domain has evolved drastically in recent years. Multiple frameworks such as Local Interpretable Model-Agnostic Explanations (LIME) [14] and SHapley Additive exPlanation (SHAP) [15] have been developed, aiming to facilitate the implementation of AI in different domains, by providing transparency and trust in underlying models. Different explainable AI solutions are already employed in the IoT domain. An IoT system [16] of low-cost

sensors incorporates an explainable AI decision support system. Another IoT system [17] makes use of an approach within the human-centric AI field for generating explanations about the knowledge learned by a neural network (in particular a multilayer perceptron) from IoT environments. We selected the LIME framework for our implementation, as it is one of the most robust and established solutions.

3. AI FOR DEVICE CLASSIFICATION

The input for classification is a scanning output – HTTP responses in a JSON format. The established text classification methods often suffer from large vector sizes and are less effective as the number of samples rises. The most effective method is a neural network [18], which learns automatically from examples, but suffers from a lack of results transparency. We are addressing this drawback using explainable AI. Another effective method for particular IoT use cases is Naive Bayes [19], which often serves as a robust method for data classification, but vectors representing an incident in Naive Bayes are larger than in the word-embedding methods of the neural network approach. However, in the case of IoT devices, we have experimentally identified that data is sparse, and the vector size is not large. A Naive Bayes method expects each feature in an HTTP response to be independent of all other features. Consequently, for the particular use case of classifying IoT devices, we suggest using Naive Bayes for text classification.

A. Features Used for Classification
We rely on features of HTTP responses suitable for the classification that have previously been developed and described in detail in [2]–[4]. These include HTTP response headers and the respective values, network name, HTML tree structure hash, body title, body keywords, SSL certificate issuer, and subject.

B. Data Sets
The primary data set consists of Internet scans of web interfaces. These scans are created by tools commonly used for Internet research – zmap and zgrab2 [20]. Both the HTTP default port 80 and the common alternative port 8080 were scanned in December 2020. Up to three redirects were followed to any port including HTTPS, in which case TLS negotiation was also saved. This toolset makes it possible to acquire research data in a uniform way, where zmap conducts Internet-wide (IPv4 only) scanning for open TCP ports in an optimised manner and hands over the identified services to zgrab2 for communicating on the HTTP application level, extracting response properties (headers, body, TLS, encountered errors), and formatting in a suitable way for further processing.

For the standard port, there were 51,118,537 elements, and for the alternative port, 8,343,898 elements. An element is a response (and appropriate redirects) corresponding to a single request that contains at least one proper HTTP response. We have augmented the elements in the data sets with additional features. The network name was looked up via the Maxmind GeoIP database. HTML tree hash, first title and body words were all generated from the response HTML body itself.

Secondary data sets (for 2018 and 2019) were utilised only to provide a comparison of the classification differences using the newly proposed Naive Bayes application, and the results are presented in Section 5. These older data sets are analysed in detail in [2].

We rely on the labelled set consisting of 171,791 elements developed in [2]. This 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, 9,561 IPCAM, 711 INFRA, 265 VOIP, 243 ICS, 218 IOT, 153 PRINTER, 4,175 UNCLEAR and 1,901 UNCATEGORIZED devices in this labelled set. Class motivations and definitions are described in detail in [2]. WEB devices are generic web sites. ICS devices serve some industrial purpose and thus might be the most impactful class. NET devices provide network connectivity to both residential and large-scale networks. IPCAM provide networked video surveillance or recording. INFRA devices provide infrastructure functionality for virtualised and related services. VOIP devices provide IP telephony services. IOT devices include all IoT and smart home products. PRINTER class consists of printers and printing servers. It is impossible to determine the class of UNCATEGORIZED devices while UNCLEAR are likely embedded devices without a clear role but not serving a WEB role.

C. Comparison with the Neural Network Classification

To classify IoT devices, we examined two AI models. The first model [2] was the Neural Network (NN) with Word Embeddings, which provided good results with high classification accuracy (87%). But the drawback of this model is that classification results are difficult to explain due to the black-box description of the NN structure. The second model is a Naive Bayes (NB) classification model, which is fast and easy to implement.

The Multinomial NB is often used for document classification problems, using the frequency of the words existing in the document as input for the calculation. The main difference between the NB model and the NN model is that the predictors are regarded as independent. Examining features extracted from IoT device responses, we concluded that they should not necessarily be regarded as dependent, because they come from different independent response parts such as the header, and different

body parts, which describe independent aspects such as the domain, company name, technologies. The position of a word within a sentence and possible relations between the words can be omitted in the case of sparse data in an IoT device response. Therefore, we can assume the independence of IoT features and employ the NB model for classification. Additionally, as described in a comparative NN vs NB study [21], NNs have a long training time and require a large number of parameters that are best determined empirically. It has been observed that the NB classifier outperforms ANN learning algorithms in all cases. NB is a generative model, which assumes conditional independence, as in the IoT devices case, whereas NNs are discriminative models, which only model the probability of the class given the input, as described in [22] and [23].

Another consideration supporting the NB model was that for meaningful NN training, long feature vectors are better, whereas IoT device feature vectors are often quite short and NB models could perform better in this particular case. The NB model is also a better match to explain it afterwards using the LIME [24] model. The LIME model explains the predictions of NB classifiers, providing rational numbers and associated features as text, which allows the human interpreter to understand if the feature word was negative or positive for each word in the IoT device response. With the LIME model, we aim to understand specific predictions to investigate the NB model whenever we doubt a given classification.

D. General IoT Device Classification Workflow

Device classification employs feature extraction and training of the NB model for 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 and 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). The workflow process is composed of two parts. One process is NB 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 using extended knowledge from new device crawls. Figure 1 provides an overview of device classification using NB. 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 execute parsing, filtering, and normalising of the content. The final classification result is based on querying the model (Step 3) or cache if the sample hash is already known and is a report in the form of a particular class name. To explain classification results, we query an explainable model (Step 4) based on the NB model and receive features with their severity for a particular class.

FIGURE 1: THE WORKFLOW FOR FEATURE EXTRACTION, DEVICE CLASSIFICATION, AND CLASSIFICATION EXPLANATION USING A NAIVE BAYES
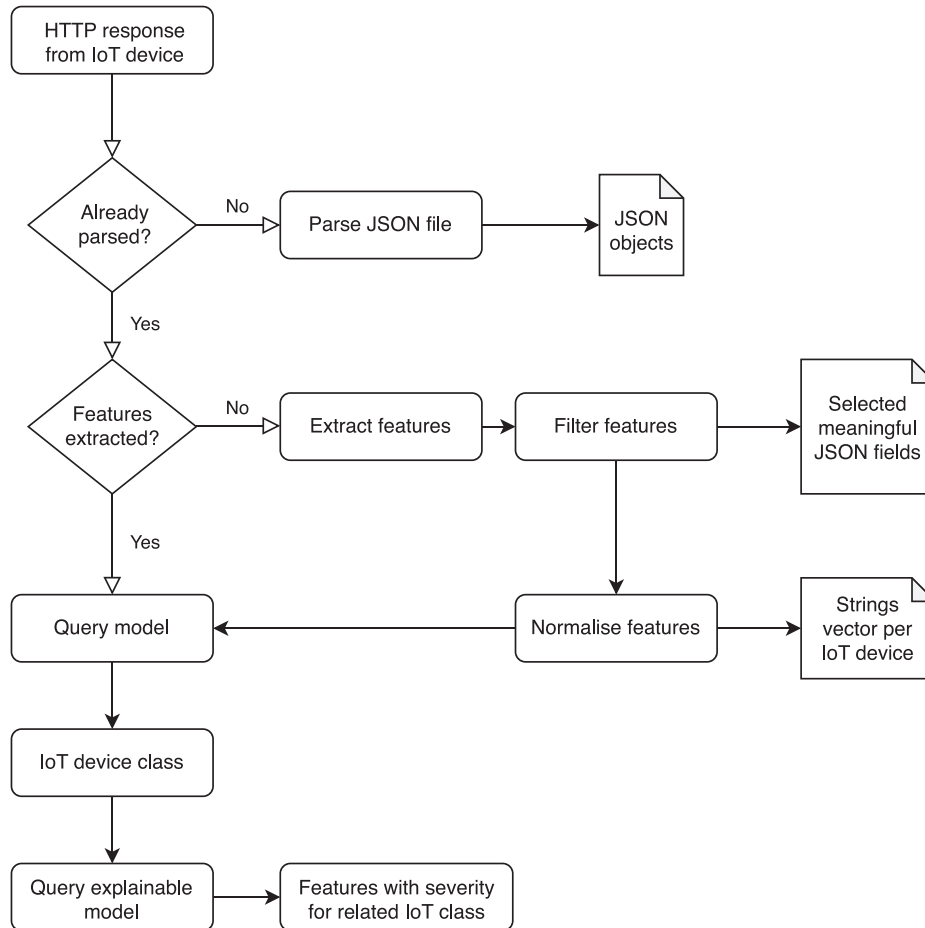
Having an HTTP response from IoT devices in the form of a JSON file, we can classify the given IoT device description to one of the earlier defined IoT device classes employing the Naive Bayes algorithm (1). This formula shows the probability of the IoT device description D (2) belonging to the IoT device class c. The probability of the IoT device description D is a product of all specifications ds that are comprised in the IoT device vocabulary.

$$p(c|D) = \frac{p(D|c)p(c)}{p(D)} \tag{1}$$

$$D = (\, ds_1, ds_2 \ldots ds_{10}\,) \tag{2}$$

The Naive Bayes algorithm picks the IoT device class with the highest probability and reports this as a classification result.

E. Naive Bayes and Explainable Model Training

The data for NB model training is prepared as described in Figure 1 in the previous section. We start with the definition of IoT classes. In the next step, we load a labelled dataset, dividing it into train, test and validation datasets. After acquisition and feature extraction, the input for the model is a list of words for each sample. For the detection of the most valuable features, we apply a TF-IDF vectoriser. TF-IDF helps to exclude words that are too frequent. For tokenising, filtering and normalising features, we filter out common and stop words, remove punctuation and special characters, remove non-alphanumeric characters, convert to lower case to have case insensitive matching, and normalise size. In the tokenising step, we break down each sentence to a set of single words. This is then converted into the one-hot vector to be processed at the input level of the NB model in Figure 2. To perform training, features aggregated in text form must be converted into numerical values, since machine learning algorithms 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 using 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 to 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. The NB training and accuracy calculation process took 15.723163 seconds. To explain classification results, we create an explainable model by creating a pipeline for the previously calculated NB model, using a vectoriser. Using the vectoriser we create a LIME text explainer for the classes defined in the first step of the workflow. Finally, using the LIME text explainer, we explain the classes predicted by the NB model and obtain related class features and weights for each query sample. The LIME text explainer calculation took 3.345 seconds. Each query takes approximately 1 second for the whole workflow.

FIGURE 2: THE WORKFLOW FOR NAIVE BAYES AND EXPLAINABLE MODEL TRAINING
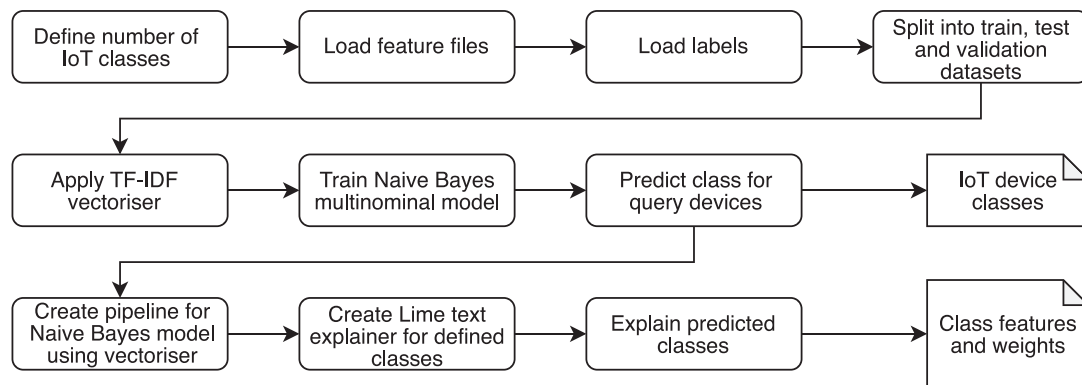
We trained a balanced model to avoid the bias in the original large labelled dataset, because we identified that by randomly sampling the classified output of the whole data set the small model performed better. As the full labelled data set primarily consists of WEB devices, the classified output is significantly skewed towards classifying devices as WEB. To avoid the 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 82%. Therefore, the classification results can be interpreted by humans able to reason and explain why a certain classification was derived. We can acquire the explanations for different features in a numerical form, meaning their weights with positive and negative signs, which means that words that are weighted negative towards one class may be positive towards another.

4. UNDERSTANDING THE CLASSIFICATION

Explainable classification can further increase the precision and also transfer the new knowledge back to experts. We review a randomly selected device from each class in an attempt to understand the classification and to evaluate options for improving it. We present the calculated prediction of classes and the most impactful weights of the features determining the likely classes.

A Cisco IP telephony device classified as VOIP is presented in Figure 3. While an expert would focus on the keywords "Cisco" and "SPA", the classifier selects "spa" as the highest weight feature and disregards "cisco" as manufacturing a large variety of NET devices. While authentication headers are more indicative of other lower power

and cheaper devices and have negative weight in this case, this is counterweighted by a slightly more complex and secure variant instead of plain text.

FIGURE 3: CLASS PREDICTIONS AND FEATURE WEIGHTS FOR A VOIP DEVICE
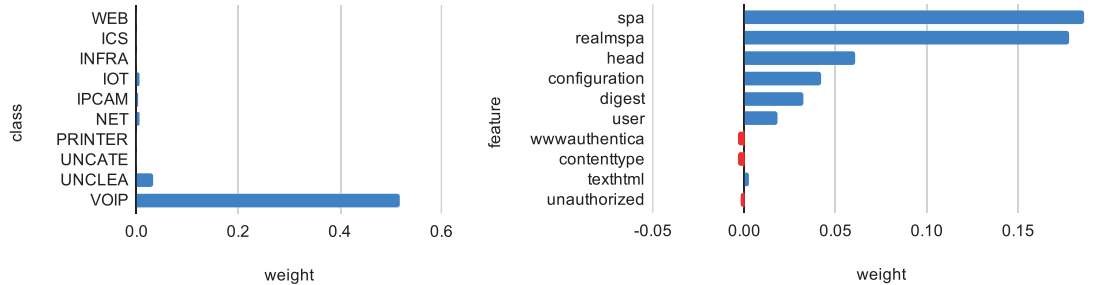
A PRINTER device is presented in Figure 4. The highest weight features "hp" and "officejet" correspond to one of the most common printer series covered by most static rule sets. The feature "broadband" is weighted negatively as is more expected in the context of networking devices. The "finance" keyword is part of the network name feature, which is not common in other randomly reviewed devices here.

FIGURE 4: CLASS PREDICTIONS AND FEATURE WEIGHTS FOR A PRINTER DEVICE
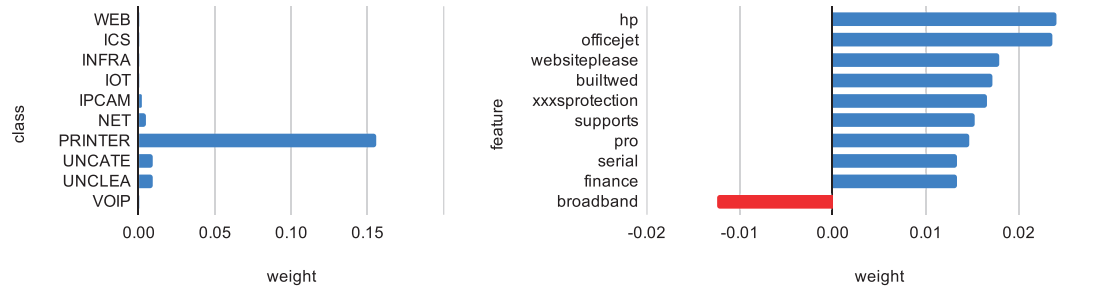
A smart home automation device from LOXONE classified as IOT is presented in Figure 5. While none of the high-weight features identifies the vendor or model, which is the way an expert would write a static rule for this device, the keyword "webinterface" for the interface and response headers has the highest weight. At the same time, security headers are weighted negatively, indicating that the model expects IOT devices to have less security features. Interestingly, a network name feature consisting of "austria" and "telekom" indicates that the manufacturer, based in Austria, has a high presence in Austrian networks. While this can be intuitively recognised by an expert, the variety of devices and complexity of the rule has prevented this from being implemented in static classification rule sets.

FIGURE 5: CLASS PREDICTIONS AND FEATURE WEIGHTS FOR AN IOT DEVICE

A VMware Horizon device classified as INFRA is presented in Figure 6. By the definition of the class, most VMware solutions match INFRA, thus the high weight of the keyword "vmware", as well as all other classes assigning a negative value to it, is unsurprising. The product keyword is also expected; static classification rule sets might contain a simple rule matching these two keywords together.

FIGURE 6: CLASS PREDICTIONS AND FEATURE WEIGHTS FOR AN INFRA DEVICE
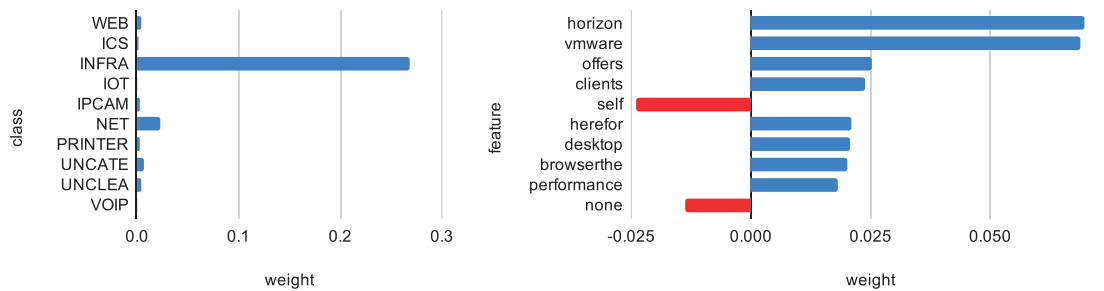
From the randomly selected devices, the spectrum analyser has the least features and is classified as ICS and presented in Figure 7. ICS devices can have the least properties of the responses that can be extracted as features. Rich response features typically weight heavily against the device being classified as an ICS. While the combination of "spectrum" and "analyser" can be evident for humans, these are treated as separate features and "spectrum" is weighted against this class while being weighted heavily in favour of some other classes. This identifies an issue with introducing network names as a feature, in this case, likely the large ISP named Spectrum, suggesting that the network name feature should be treated differently from the response features.
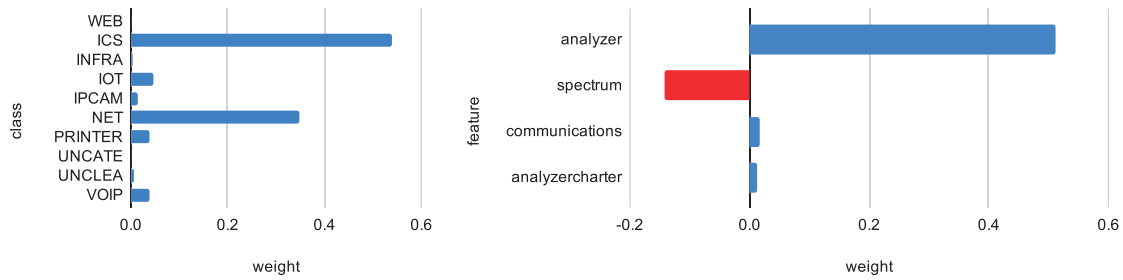
For the network device presented in Figure 8 (NET class), the second highest weight feature "gateway" is a classic keyword even in static rules. The feature set and raw response confirm that this is an unbranded residential network gateway for which even an expert is unable to extract more information without active probing. This feature is not unique to the NET class. It might correspond to gateway functionality in an application protocol sense or display configuration debugging information for any networked device. In this particular case, this feature is weighted in favour of only the VOIP class. Most of the remaining determining features consist of authentication interface keywords, including the highest weight feature "incorrect" indicating failed authentication. The way an authentication interface is presented has a high weight in determining the class.

FIGURE 8: CLASS PREDICTIONS AND FEATURE WEIGHTS FOR A NET DEVICE
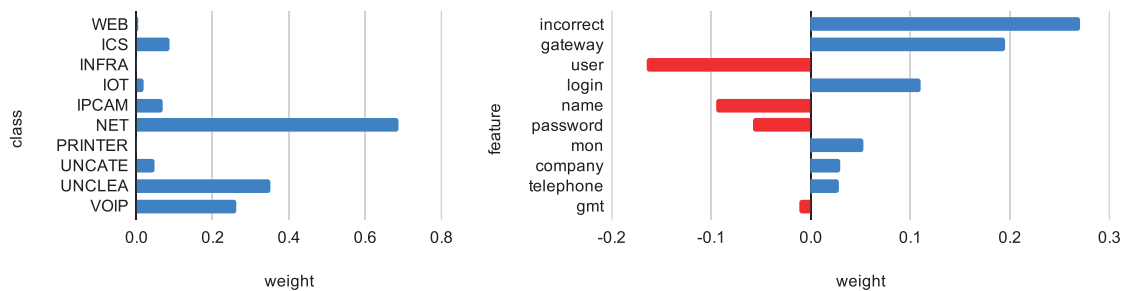
A Hikvision networked surveillance device classified as IPCAM is presented in Figure 9. The "dnvrswebs" is a software version unique to IP cameras and video recorders and thus is weighted heavily. In general, it is weighted negatively against all other classes. Most static rule sets have this as a simple match rule to reliably classify IP cameras.

FIGURE 9: CLASS PREDICTIONS AND FEATURE WEIGHTS FOR AN IPCAM DEVICE
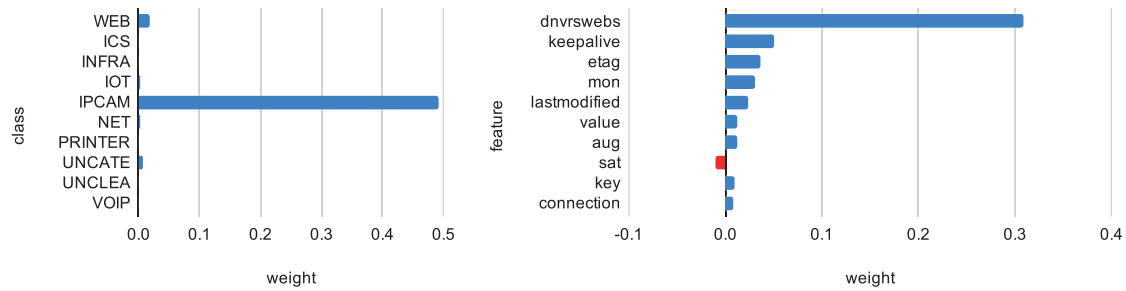
The WEB device presented in Figure 10 is an Apache Tomcat interface allowing the deployment and management of web applications. While the feature "apache" has a high weight in determining the class, it is not always the case, otherwise a blanket static rule would suffice. In general, it has a negative weight on the UNCLEAR class where no web sites are expected. The keyword "restricted", generally associated with web interface authentication, has a significant negative weight.

FIGURE 10: CLASS PREDICTIONS AND FEATURE WEIGHTS FOR A WEB DEVICE
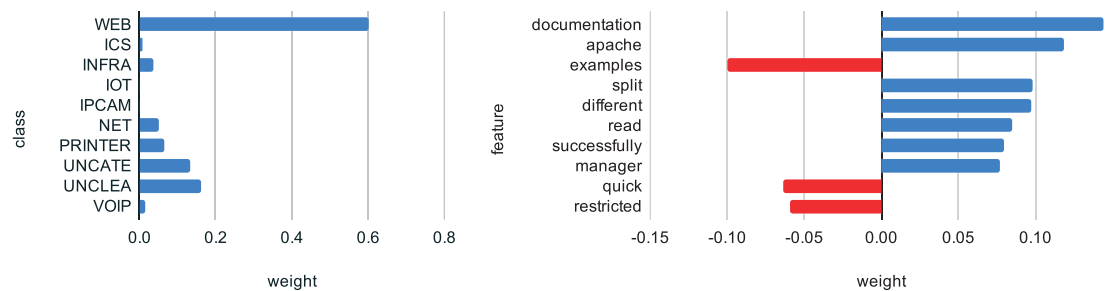
A device classified as UNCLEAR and likely an embedded device without a determinable functionality but definitely not a generic web site is presented in Figure 11. Keywords related to HTTP basic authentication and the displayed message are weighting in favour of this class; while the presence of the Server header revealing software name and version is weighting against, as it is often a high-weight feature, in this case, it is a generic embedded software having many uses.

FIGURE 11: CLASS PREDICTIONS AND FEATURE WEIGHTS FOR AN UNCLEAR DEVICE
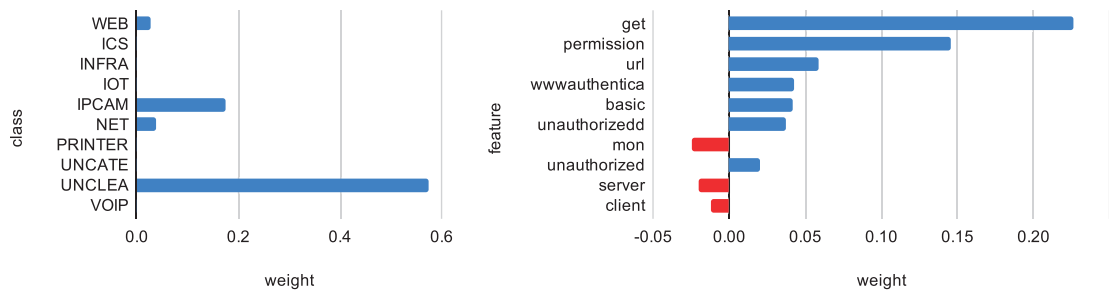
An UNCATEGORIZED device that cannot be determined as part of any class is presented in Figure 12. This device has a small feature set (all generic) but not as small as the most basic embedded devices. The generic response headers are sufficient to be also those of a website or service not handling default requests. In general, plain text content type, which is the heaviest weighted feature, corresponds to an unformatted output of mostly short error messages. From this set of features, an expert is not able to reliably determine the class either.

FIGURE 12: CLASS PREDICTIONS AND FEATURE WEIGHTS FOR AN UNCATEGORIZED DEVICE
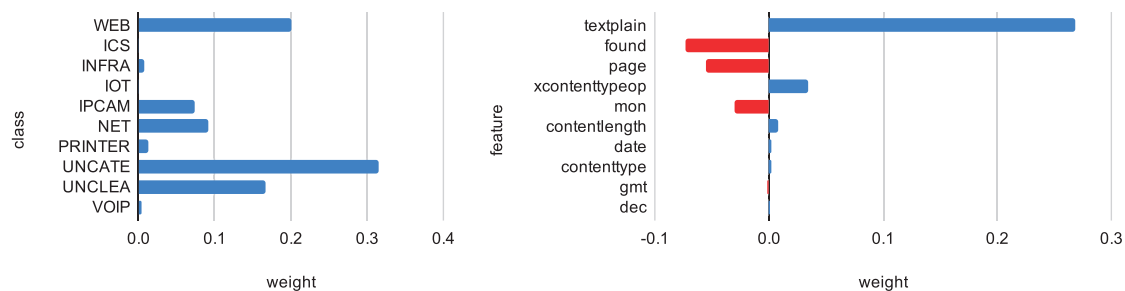
While there can be identified cases that are common and covered by static classification rule sets even within these few random examples, a more complex classification matching expert intuition can be seen. These types of cases can be classified individually by an expert, but defining all of that into static rules is not feasible, not only because of the sheer number of rules but also the complexity which would require statistical calculations to formalise the intuition.

5. CLASSIFICATION RESULTS

The relative class distribution is presented in Figure 13; the Naive Bayes classifier results developed in this paper are prefixed NB. The remaining classification data are based on neural network results from [2]; the raw scan data from the same source is used to test the Naive Bayes classification for 2018 and 2019, while the 2020 data set has been created specifically for this research.

While classification differences can be easily observed, they are explained by the varying accuracy ranges between different methods. Although the goal of this research was not to analyse the classification, we can identify the main trends in Naive Bayes classification. An increase in ICS devices is unexpected in light of worldwide efforts to disconnect these devices from the Internet; most likely these are new deployments of low impact automation devices. The decrease of INFRA devices is expected with a shorter lifecycle of deployments and new deployments following better security

practices. The stable proportion of IOT devices is positive, considering the increasing number of new deployments. IP cameras, which often require remote reachability, see a slight increase, and by contrast, NET devices, which do not, see a significant decrease. IP telephony devices experience a stable decrease.

FIGURE 13: THE NEW NB CLASSIFICATION APPLIED FOR 2018–2019 AND NEURAL NETWORK CLASSIFICATION
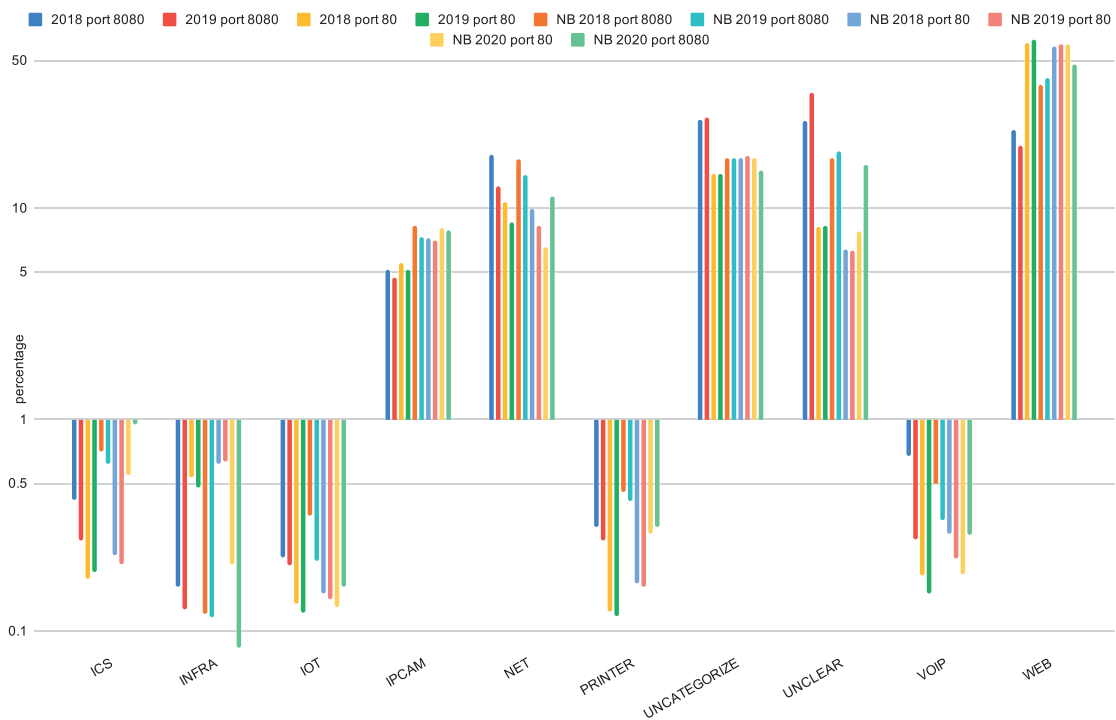
6. CONCLUSIONS AND DISCUSSION

Device classification is an important emerging research field. While existing neural network classifiers already provide classification with high levels of accuracy [2], [9], the results are not always understandable by human experts. In some of these cases, it is hard to distinguish who is in the right. An expert often has the ability to validate his predictions through active probing and other external sources of intelligence. But what if the device is not present on the Internet anymore? This often happens because of the dynamically assigned IP address change, a device going offline or when analysing historical data sets. The expert is left with only the feature set and potentially the raw data from which features were extracted to make a decision. Features are numerous and while clues could be found and even validated using external knowledge, there is no confirmation that these were decisive features in the classification, so no correction in the classifier can be made.

Understanding the classification brings result transparency – the ability to explain the predicted class. With our suggested combined Naive Bayes and LIME approach, we were able to demonstrate a reliable method for the explainable classification of devices with a web interface being reachable on the Internet. Understanding the features used by the model for class prediction permits better analysis of the device properties and, consequently, improvements in the classification accuracy via a more targeted handling of device data and feature filtering. This approach supports a better general understanding of higher risk potentially vulnerable devices on the Internet and, subsequently, can increase not only security for the device owner but also overall security.

REFERENCES

[1]   Y. Jia, B. Han, Q. Li, H. Li, and L. Sun, "Who owns Internet of Thing devices?," *International Journal of Distributed Sensor Networks*, vol. 14, no. 11, Nov. 2018, doi: 10.1177/1550147718811099.

[2]   A. Lavrenovs, R. Graf, and K. Heinaaro, "Towards Classifying Devices on the Internet Using Artificial Intelligence," in *2020 12th International Conference on Cyber Conflict (CyCon)*, Estonia, May 2020, pp. 309–325, doi: 10.23919/CyCon49761.2020.9131713.

[3]   A. Lavrenovs 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: 10.1109/TELFOR48224.2019.8971100.

[4]   A. Lavrenovs 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.

[5]   A. Mirian *et al.*, "An Internet-wide view of ICS devices," in *Proceedings of 2016 14th Annual Conference on Privacy, Security and Trust (PST)*, Dec. 2016, pp. 96–103, doi: 10.1109/PST.2016.7906943.

[6]   M. Dahlmanns, J. Lohmöller, I. B. Fink, J. Pennekamp, K. Wehrle, and M. Henze, "Easing the Conscience with OPC UA: An Internet-Wide Study on Insecure Deployments," in *Proceedings of the ACM Internet Measurement Conference*, Virtual Event USA, Oct. 2020, pp. 101–110, doi: 10.1145/3419394.3423666.

[7]   X. Feng, Q. Li, H. Wang, and L. Sun, "Acquisitional Rule-based Engine for Discovering Internet-of-Things Devices," in *27th USENIX Security Symposium (USENIX Security 18)*, Baltimore, MD, Aug. 2018, pp. 327–341, [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/feng

[8]   M. Nawrocki, T. C. Schmidt, and M. Wahlisch, "Uncovering Vulnerable Industrial Control Systems from the Internet Core," in *NOMS 2020 - 2020 IEEE/IFIP Network Operations and Management Symposium*, Budapest, Hungary, Apr. 2020, pp. 1–9, doi: 10.1109/NOMS47738.2020.9110256.

[9]   K. Yang, Q. Li, and L. Sun, "Towards automatic fingerprinting of IoT devices in the cyberspace," *Comput. Netw.*, vol. 148, pp. 318–327, Jan. 2019, doi: 10.1016/j.comnet.2018.11.013.

[10] Y. Meidan *et al.*, "ProfilIoT: a machine learning approach for IoT device identification based on network traffic analysis," in *Proceedings of the Symposium on Applied Computing*, Marrakech Morocco, Apr. 2017, pp. 506–509, doi: 10.1145/3019612.3019878.

[11] A. Sivanathan *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.

[12] B. Bezawada, 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.

[13] P. Yadav, A. Feraudo, B. Arief, S. F. Shahandashti, and V. G. Vassilakis, "Position paper: A systematic framework for categorising IoT device fingerprinting mechanisms," in *Proceedings of the 2nd International Workshop on Challenges in Artificial Intelligence and Machine Learning for Internet of Things*, Virtual Event Japan, Nov. 2020, pp. 62–68, doi: 10.1145/3417313.3429384.

[14] A. Ignatiev, "Towards Trustable Explainable AI," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Yokohama, Japan, Jul. 2020, pp. 5154–5158, doi: 10.24963/ijcai.2020/726.

[15]   S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017, vol. 30, pp. 4765–4774, [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[16]   N. L. Tsakiridis *et al*., "Versatile Internet of Things for Agriculture: An eXplainable AI Approach," in *Artificial Intelligence Applications and Innovations*, vol. 584, I. Maglogiannis, L. Iliadis, and E. Pimenidis, Eds. Cham: Springer International Publishing, 2020, pp. 180–191.

[17]   I. Garcia-Magarino, R. Muttukrishnan, and J. Lloret, "Human-Centric AI for Trustworthy IoT Systems With Explainable Multilayer Perceptrons," *IEEE Access*, vol. 7, pp. 125562–125574, 2019, doi: 10.1109/ACCESS.2019.2937521.

[18]   S. L. Y. Lam 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.

[19]   C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. New York: Cambridge University Press, 2008.

[20]   The ZMap Team, "The ZMap Project." https://zmap.io/ (accessed Aug. 30, 2017).

[21]   Md. S. Islam, S. M. Khaled, K. Farhan, Md. A. Rahman, and J. Rahman, "Modeling Spammer Behavior: Naive Bayes vs. Artificial Neural Networks," in *2009 International Conference on Information and Multimedia Technology*, Jeju Island, Korea (South), 2009, pp. 52–55, doi: 10.1109/ICIMT.2009.48.

[22]   T. M. Mitchell, "Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression," in *Machine Learning*, 2nd-draft ed., 2017.

[23]   X. Daniela, C. Hinde, and R. Stone, "Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," *International Journal of Computer Science Issues*, vol. 4, 2009.

[24]   M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA, Aug. 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

Towards an AI-powered Player in Cyber Defence Exercises

Roland Meier
Department of Information Technology and Electrical Engineering
ETH Zürich
Zürich, Switzerland
meierrol@ethz.ch

Artūrs Lavrenovs
NATO CCDCOE
Tallinn, Estonia
arturs.lavrenovs@ccdcoe.org

Kimmo Heinäaro
NATO CCDCOE
Tallinn, Estonia
kimmo.heinaaro@mil.fi

Luca Gambazzi
Science and Technology
armasuisse
Thun, Switzerland
luca.gambazzi@armasuisse.ch

Vincent Lenders
Science and Technology
armasuisse
Thun, Switzerland
vincent.lenders@armasuisse.ch

Abstract: Cyber attacks are becoming increasingly frequent, sophisticated, and stealthy. This makes it harder for cyber defence teams to keep up, forcing them to automate their defence capabilities in order to improve their reactivity and efficiency. Therefore, we propose a fully automated cyber defence framework that no longer needs support from humans to detect and mitigate attacks within a complex infrastructure. We design our framework based on a real-world case – Locked Shields – the world's largest cyber defence exercise. In this exercise, teams have to defend their networked infrastructure against attacks, while maintaining operational services for their users. Our framework architecture connects various cyber sensors with network, device, application, and user actuators through an artificial intelligence (AI)-powered automated team in order to dynamically secure the cyber environment. To the best of our knowledge, our framework is the first attempt towards a fully automated cyber defence team that aims at protecting complex environments from sophisticated attacks.

Keywords: *artificial intelligence, automation, Locked Shields, cyber defence, security*

1. INTRODUCTION

Attackers and defenders of cyber systems often engage in an arms race on many levels: attacks become more sophisticated, happen at a faster pace, use greater stealth, and show less evidence. Defenders often lag behind attackers because of the underlying asymmetry between them needing to defend a heterogeneous infrastructure against all possible attack vectors and the attackers only needing to find a single or limited number of vulnerabilities to exploit.

It is widely accepted that a human workforce alone cannot keep up with the workload in a typical cyber environment. As a consequence, many tools have been developed to support human defenders: from pattern detection tools [1], [2] to sophisticated anomaly detection using machine learning [3]–[5]. However, while these tools significantly reduce the effort required by humans to detect anomalies, they do not automate the entire defence process and still require significant human labour to mitigate the impact of attacks and keep services running.

In this paper, we develop a novel framework for fully automated cyber defence. Our framework is designed for Locked Shields (LS) [6], [7], the world's largest live-fire cyber defence exercise but is expected to be applicable to the defence of many cyber infrastructures. In LS, human teams from different nations have to defend their networked infrastructure against a series of attacks while maintaining operational services for several days. In contrast to classical defence teams in LS, which may consist of large numbers of human players, our framework consists of an AI-powered system without any human intervention once the exercise has started.

To the best of our knowledge, our framework is the first attempt towards a fully automated cyber defence team. In particular, our contributions in this paper are:

- A description of the situation during LS for attackers and defenders (Section 2);
- A framework architecture for an automated team to participate in LS (Section 3);
- System descriptions of the main components of this framework including sensors and situational awareness (Section 4), actuators (Section 5), AI engine (Section 6) and control logic (Section 7); and
- A case study highlighting how our framework is able to mitigate the impact of attacks in LS (Section 8).

Related Work: A cyber exercise such as LS can be considered as a game with complex and continuously changing rules. Past developments in AI have resulted in computers

outperforming human players in games such as chess and go. Game algorithms have traditionally been designed for a single game with predefined rules. However, the authors of [8] also apply AI for learning to play chess without a priori knowing the rules. AI adapting to the rules of several games to achieve a general game system has also been studied in [9]. The problem we consider in this work is different as the attacker does not have to follow a predefined set of moves or rules.

The concept of cyber defence exercises and their importance in training experts and testing defence processes has been studied in [10]. The use of AI in defending IT systems has been considered beneficial for years. Already in 2011, Tyugu described that defending cyberspace cannot be handled by humans due to the speed of processes and the amount of data involved [11]. Using machine learning to detect malicious traffic in industrial networks has been studied in [12]. Bogatinov et al. predict that autonomous AI agents defending IT systems will gradually become better than humans [13]. On the other hand, there is a constant arms race between defenders and attackers also in cyberspace. If AI can be used to improve cyber defence, it can also be used to improve attacks and malware as is studied in [14] and [15].

2. BACKGROUND ON LOCKED SHIELDS

Below, we summarise key aspects of LS: the roles of the different teams, and the scoring system.

A. Teams and Organisation

The two most important teams concerning this paper are the defenders (blue team) and the attackers (red team), which we explain in more detail below. Table I provides an overview of all teams.

TABLE I: TEAMS IN LS

Team name	Role	Explanation
Blue Team (BT)	Defender	Needs to defend its infrastructure against attacks from the RT while maintaining availability.
Red Team (RT)	Attacker	Executes attacks against the infrastructure of each participating Blue Team.
Green Team (GT)	Infrastructure operator	Creates and maintains the technical exercise infrastructure.
Yellow Team (YT)	Monitoring	Provides situational awareness during the exercise.
User Simulation Team (UST)	Benign users	Legitimate system users with poor cyber hygiene conducting activities on BT systems.
White Team (WT)	Organiser	Responsible for non-technical aspects of the exercise.

Blue Teams (BTs) are the main training audience. There are more than 20 BTs and each of them acts independently within its dedicated infrastructure (the Gamenet). The main tasks of each BT are to harden the provided Gamenet to be available and resilient before the attacks start and to defend them against ongoing attacks.

The **Red Team** (RT) conducts attacks to accomplish predefined objectives (e.g. gain control of a device). The RT is not the training audience, and so its activities are highly regulated to maximise fairness. The RT has three sub-teams: Web, Client-side (CS), and Network and special systems (NET + SS). SS consists of all ICS and cyber-physical systems. The RT knows vulnerabilities in advance and uses a large toolset to attack. It includes a command and control (C&C) infrastructure, custom, and known malware. In addition, the RT takes advantage of user(s) to run commands (e.g. download malware) on Gamenet machines.

BT communication channels with other teams: Besides defending their systems, the BT is required to maintain predefined communication channels as listed in Table II.

TABLE II: COMMUNICATION CHANNELS

With	Channel	Purpose
GT	Web ticketing, chat	Request manual reverting of SS, receive or request information about Gamenet status
YT	Web interface	Provide periodical or on-demand reports, threat reports, key events, situation reports, adversary assessments
UST	Web ticketing	Read, address and respond to UST tickets
WT	Voice or video call	Verify voice or video call functionality in the Gamenet, voice or video reporting

B. Scoring

LS has a complex scoring system that changes yearly, but the most important high-level objectives from a BT perspective remain constant: (i) prevent the RT from reaching its objectives; (ii) keep systems available; and (iii) interact with the UST and YT.

RT objectives: The RT follows a list of objectives it wants to achieve in each phase of the gameplay. After each phase, the scoring is updated based on the achieved objectives in each Gamenet.

Availability: All systems in the Gamenet need to be available as specified in the rules. Mistakes by the BT (e.g. over-aggressive firewall rules) or attacks from the RT (e.g. compromised services) reduce the availability score.

Interactions with UST and YT: The UST continuously accesses systems and services in the Gamenet. If a service does not work as expected, it creates a support ticket and the BT needs to solve the issue. Otherwise, it will incur a scoring penalty. In addition, the YT periodically requests reports about failed and successful attacks from BTs.

3. ABT FRAMEWORK OVERVIEW

We now provide an overview of our framework to implement an automated BT (ABT). Our framework is designed for LS, but it could easily be adapted to other environments. In our framework, the skilled human players in a traditional BT are replaced with an architecture as shown in Figure 1.

FIGURE 1: ARCHITECTURE OVERVIEW
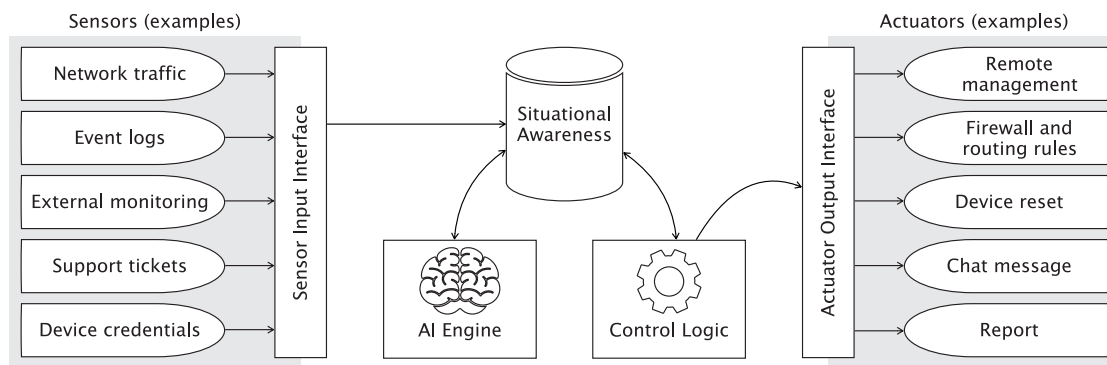
There are five types of building blocks:

- **Sensors** provide measurements and data;
- **Actuators** trigger actions;
- The **situational awareness database** contains all the sensor data, the AI state and external data (e.g. logs, AI models and malware signatures);
- The **AI engine** continuously (re-)learns models (e.g. to detect anomalies);
- The **control logic** determines actions for actuators to execute based on the available information in the situational awareness database.

4. SENSORS AND SITUATIONAL AWARENESS

We now describe how the ABT maintains situational awareness by collecting data and building models based on these data. We group the sensors in five categories: application-level, device-level, network-level, user-interaction, and organisational inputs.

A. Application Sensors

Application state and configuration: This includes for example the contents of the database, the registry or the filesystem. This allows the ABT to identify misconfigurations, vulnerable components, weak credentials, unauthorised modifications, and to restore a compromised application. The application state can be retrieved at the beginning of the exercises (e.g. for static configurations) or periodically (e.g. for dynamic state).

B. Device Sensors

System information and logs: For example, process creations, filesystem activities, patch level, privileged activities, machine fingerprinting, interaction with remote services, and domain controllers. Besides logs that are collected by default, the ABT may log additional events and forward them to the situational awareness database.

Remotely executed command outputs: The ABT evaluates and collects the outputs of commands executed by the actuators on Gamenet devices.

Active monitoring: The ABT deploys sensors to perform on-demand device monitoring or active probing. This could not be possible on some devices (e.g. SS) because of resource constraints or because of their proprietary architecture. Monitoring agents on workstations can allow logging mouse and keyboard inputs interacting with proprietary GUI applications.

C. Network Sensors

Network traffic: By default, one virtual machine (VM) receives a real-time stream of packets captured in the Gamenet. The traffic is captured at particular devices in the network (switches and routers). Therefore, it only contains packets crossing these devices and can contain the same packet multiple times. To capture additional traffic, the ABT configures additional sensors, for example by recording traffic at hosts.

Network configuration: To discover the network topology and configuration, the ABT relies on network scanning, management and diagnostic tools such as nmap, snmp, or traceroute in addition to the organisational inputs (see below). These sensors

allow it to determine the actual behaviour of devices and compare this with the specifications.

D. User-Interaction Sensors

Support tickets: The UST creates support tickets that indicate problems in the Gamenet (usually related to availability and functionality). This input is important for situational awareness because functionality issues are especially hard to detect automatically.

E. Organisational Inputs

System and credential list: The ABT receives a list of all systems in the Gamenet and credentials to access them with administrative privileges. The ABT leverages available remote access services (e.g. SSH, telnet, RDP, and web interfaces) for situational awareness.

Scoring system and external monitoring: The primary components of the scoring are objectives accomplished by the RT and the availability state of all systems. These inputs are essential for situational awareness to understand how successful or unsuccessful actions and attack mitigations have been in near real-time.

F. Situational Awareness Database

The situational awareness database contains all the information that is relevant for the ABT to succeed. It is a central entity that includes all collected sensor data, any learned models from the AI engine, actuator output as issued by the control logic, as well as relevant information from outside the context of LS, such as known software vulnerabilities or indicators of compromise. The collected and processed data is stored persistently and made available to the AI engine and the control logic. The ABT represents the stored data in a global knowledge graph in order to reason about the current situation including the Gamenet and the previous interactions with all the other teams.

The data sources are heterogeneous and the data must first be pre-processed and then fused in order to allow queries over data coming from multiple sources and formats. For example, a query to the database might correlate data from device and network sensors in order to identify compromised services that leave attacker traces in both data sources. Furthermore, it is possible to correlate system failures identified from support tickets with application state, so that the control logic can derive the services that may need a restart or a recovery.

5. ACTUATORS

We will now describe the actuators at the application, device and network level, as well as actuators for interacting with humans from other teams.

A. Application-level Actuators

Web application firewall configuration: To block the common attack vectors (e.g. malicious file uploads), the control logic deploys policies to inspect requests and to block those which – according to the AI models – are malicious.

Application configuration and patch: Change the configuration of an application, patch vulnerable libraries, or update credentials.

B. Device-level Actuators

Remote management: This actuator initially allows the ABT to log in to any device with administrative privileges. If possible, the ABT installs tools to enable effective remote management and infrastructure as a code (IaaC). Having multiple options for remote access improves the ABT's chances of regaining access after a system was compromised and allows consistent management of operations on the Gamenet.

Restoring a device's initial state: The ABT can revert a device to its original state (resulting in a scoring penalty). VMs can be reverted automatically by the ABT's actuators, while some special systems require manual intervention by the GT.

C. Network-level Actuators

Device configuration and rules: To control network traffic, the ABT uses actuators to dynamically adapt the configuration and the rules of firewalls and routers. Initially, the devices are set up as simply as possible: firewalls let everything pass and routers merely guarantee connectivity. The ABT configures these devices to limit authorised traffic, and – if possible – replaces the software running on them. The actuator can also apply a new rule to block traffic from or for compromised devices. To gain higher control over the network traffic, the actuator may also change the routing behaviour of the devices; for example, to configure the network such that all traffic passes a programmable controller.

D. User-Interaction Actuators

Chat and email messages: The ABT has to address and respond to messages sent by other teams. This actuator thus generates human-readable chat and email messages based on the current situation.

Incident reports: The ABT also needs to provide reports about successful and failed attacks to the GT. This actuator reports those incidents to the GT.

6. AI ENGINE

The AI engine's task is to learn models from data and infer facts from historical events in the Gamenet. It gets its data from the situational awareness database, performs machine learning techniques, and feeds the learned models and facts back to the database. Below, we present a categorisation of AI approaches and then sketch how models are trained.

A. AI Categorisation

We characterise each AI-enabled defence according to three dimensions:

1. Level: How sophisticated is the AI? Example: Narrow AI;
2. Tasks: What does the AI do? Example: Classify samples;
3. Type: Which AI technique is used? Example: unsupervised learning.

The four levels of AI:

- **Level 0 – Reactive narrow AI:** Level-0 AI is very basic in the sense that it only reacts to current inputs. This restricts it to simple decisions (e.g. if a network packet has destination port 22, drop the packet). This level of automation is the same as many existing tools (e.g. [1], [2], [16], [17]). For example, popular intrusion detection systems, such as Snort [2], check network traffic for known signatures, or antivirus software checks executables against a database of known malware. The contribution of Level 0 should not be underestimated in our context. Using external sources (rules, IOC, patterns, etc.), it is possible to identify traces of attacks with few computational resources.
- **Level 1 – Limited-memory narrow AI / Weak AI:** Level-1 AI is what is typically understood as machine learning today (e.g. Siri, Watson, self-driving cars). Many proposed machine learning solutions [18]–[26] already solve tasks that are important for the ABT. This kind of AI is equipped with data storage and learning capabilities that enable the ABT to use historical data to make informed decisions.
- **Level 2 – General AI / Strong AI / Deep AI:** Level-2 AI mimics human-level intelligence but it does not exist yet. We, therefore, do not consider it in this work.

- **Level 3 – Super AI:** Level-3 AI is considered self-aware AI that surpasses human intelligence. It is not entirely clear if this kind of AI can be achieved and is also out of the scope of our work.

The five tasks for AI we consider useful are:

- **Identification/classification:** tell me what the thing is, such as detecting command and control flows (e.g. [18]); intrusion/anomaly detection (e.g. [19]–[21]);
- **Categorisation:** group similar things together such as log clustering for detecting security issues (e.g. [22]);
- **Assessment:** tell me whether I should care about this thing, for example, to prioritise security events (e.g. [23]);
- **Recommendation:** tell me what to do about this thing in order to, for example, recommend defence actions (e.g. [24]);
- **Prediction:** tell me if this thing, for example, predicts attacks or vulnerabilities (e.g. [25] and [26]).

Three types of AI exist:

- **Supervised:** learning from labelled training data. Supervised approaches can use the situational awareness data from the current and/or past iterations of LS for training;
- **Unsupervised:** detecting previously undetected patterns in unlabelled data, using, for example, clustering techniques;
- **Reinforcement:** AI takes decisions and receives feedback which it uses for future decisions.

B. AI Models for Improved Situational Awareness

Below, we sketch models and applications of the AI engine to improve the situational awareness of the ABT.

Anomaly detection: The AI engine learns the behaviour of applications, devices, and network traffic from sensor data in order to build models of normal behaviour and detect anomalies or intrusions (e.g. an application suddenly deleting many files). For unsupervised learning, the models are learned based on the actual data, while supervised learning techniques require labelled datasets from previous exercises.

Data fusion: The situational awareness database contains information from many different sources and the AI engine fuses this information to extract new insights. For example, a successful defacement attack resulting in support tickets, event log entries,

and network connections that seem unsuspicious at first is inferred as an attack by means of time correlation.

Prioritise defence actions: Often, the ABT has multiple options to restore a system after a successful attack (e.g. patching or reverting the device), but the options have different costs and side-effects (e.g. scoring penalty, system availability). The AI engine therefore models and suggests the best strategy given the information about the current situation, an objective function (e.g. the scoring rules), or potential previous observations (reinforcement learning).

Predict future attacks: Since the attacks are similar in each iteration of LS, the AI engine learns using data from previous iterations to predict which attacks will happen. Furthermore, predictive models are learned to infer if an observed attack is likely to be performed against other devices.

Interact with humans: The ABT is required to interact with humans in other teams. For this, a chatbot based on AI and natural language processing techniques (e.g. [27]–[30]) is used. The chatbot is able to ask humans about, for example, the classification of the problem and the target IP address. This information is inserted into the situational awareness database. If the chatbot is not able to understand or resolve the problem, the chatbot can ask the user for more information or try to please her with generic statements to optimise scoring points.

Summary: In Table III, we summarise the tasks of the AI engine at the different stages of the exercise categorised according to their function.

TABLE III: TASKS FOR AI MODELS

	Task				
Stage	**Identification**	**Categorisation**	**Assessment**	**Recommendation**	**Prediction**
Initial hardening	Detect misconfigurations between similar clients	Find groups of similar devices	Identify potentially vulnerable devices	Generate secure configurations	Predict events that could indicate a compromise
Monitoring & Response	Detect malicious applications/ commands/ network traffic	Detect malicious patterns in log files	Prioritise security events	Select defence / restore action	Predict future attacks
Reporting	Understand support tickets	Link support tickets and monitoring alerts	Prioritise tickets	Formulate response to ticket	Predict impact on the scoring
Recovery	Find devices that need to be recovered	Find a similar system as a template for the recovery	Determine whether a system needs to be recovered	Select recovery strategy	Predict impact on the scoring

7. CONTROL LOGIC

We will now describe the tasks managed by the control logic, first in the initialisation phase, and then in the gameplay phase. The control logic makes decisions about whether and which actuators to actuate in which way depending on the information available in the situational awareness database.

A. Defence Techniques and Goals

The goal is to defend against complex cyber attacks by hardening the system and maintaining/restoring service availability in order not to lose scoring points. Table IV provides an overview of the employed defence techniques of the control logic for the different stages before and during the exercise.

TABLE IV: DEFENCE TECHNIQUES

Stage	WEB	CS	NET + SS
Initial hardening	Set up WAF. Replace applications with secure clones. Security evaluation.	Set up a centralised management tool. Deploy antivirus, firewall, execution policies, monitoring agents.	Deploy firewall rules for all networks. Deploy network IDS/IPS. Secure interfaces (e.g. of the ICS devices).
	Identify vulnerabilities, patching software and fix misconfigurations. Change credentials for OS accounts, services and application accounts. Remove unrequired accounts and services. Identify and remove RT implants.		
Monitoring & Response	If exploited functionality in the web application can be identified after or mid-attack, disable access to it and attempt fixing.	Process creation and file access have to be monitored to identify and block suspicious activities. Monitoring tools must also be closely guarded.	Locate and remove rogue network connections and rogue hosts. If SS is misbehaving, attempt to identify the cause and deploy rules to minimise unauthorised access.
	Anomaly detection. If a suspected attack, data exfiltration or C&C channels or processes can be identified – block the relevant network traffic and terminate processes.		
Reporting	Report successful and failed attacks, provide adversary assessment.		
Recovery	Self-recovery generally possible by preparing database and source code (e.g. scripts) dumps in advance.	Self-recovery generally not possible, the automated reverting interface is available to the BT.	Self-recovery is not possible and GT has to take action.

B. Initial Hardening

Since the Gamenet initially comes in a relatively unprotected state, the first task of the control logic is to harden the devices, networks and applications based on templates of secure configurations prepared by humans.

General: The control logic replaces all credentials and disables unused services on all machines. If required, it sends the new credentials to the UST. Operating systems and software updates are installed. Wherever possible additional security software is installed (e.g. antivirus solution, monitoring agents).

Workstations: CS consists primarily of Windows workstations and domain controllers and some Linux or macOS workstations [31], [32]. Workstations are managed in a centralised manner by utilising group policies. Applying group execution policies allows us to whitelist known benign applications while blocking unknown services. Applying group application firewall policies further allows us to restrict the communications of malicious software that has bypassed execution policies. Initial policies are defined by learning workstation behaviour at the AI engine before the attack stage.

ICS devices: Defence involves protecting VMs running engineering workstations and special software combined with physical components. ICSs require initial configuration to mitigate basic vulnerabilities. Depending on the services required from the device, the ABT disables web interfaces or protects them with strong passwords.

Network: The ABT knows which services are running and need to be available at each of the devices from the organisational inputs, and it can derive firewall rules which only allow expected traffic. This is especially important in ICS protocols with no encryption or authentication. If the existing firewalls in the Gamenet are not enough to perform this action (e.g. because not all traffic crosses them), the control logic attempts to change the routing behaviour to forward the traffic through a central controller or a gateway which sees all traffic and can control it (e.g. by modifying or blocking the traffic).

Web applications: Since most of the web traffic is encrypted, it is difficult to filter on the network or transport layer and a Web Application Firewall (WAF) becomes the primary defence tool. While a WAF might only protect against basic attack vectors by default, the AI engine creates a behaviour profile based on data from the monitoring sensors and UST interactions in order to enable more aggressive filtering while maintaining functionality. In addition, the AI engine may also reveal vulnerabilities to the control logic from analysis of the web application source code.

Files: While full-disc imaging or filesystem copying is not possible per exercise limitations, calculating hashes of the file system tree on all machines is beneficial for situational awareness. When standard OS file and software hashes are available, pre-planted backdoors and non-default configurations can be detected at this stage

already. For some applications (primarily web), it is possible to create snapshots by copying files and creating database dumps.

C. Actions during Gameplay

During the LS exercise, the control logic attempts to identify attacks and mitigate their impact.

Revert and reboot: Quick recovery to a working state can often maintain the availability score. In simple cases, a reboot restores the functionality of a system. If an application can be restored from snapshots created by the ABT in the initial hardening phase, this is preferred as it does not incur a scoring penalty. If this fails, the penalised in-game revert of VMs is executed (restored state could include vulnerabilities). More complex systems might require a pre-defined order of reverting and rebooting multiple targets. SS might require separate interfacing with the GT to request manual reverting of physical devices.

Block malicious traffic: In addition to the static firewall rules, the control logic also relies on learned AI models to identify traffic that is associated with malicious activities and blocks it.

Identify senders of malicious traffic: If the sender or the receiver of malicious traffic is a device under the ABT's control, the control logic takes further actions to patch the device or block it completely.

Block malicious processes: The execution policies might not stop all malicious processes. Based on each new process' properties and event logs (or events for existing ones), the control logic uses learned AI models to determine if the process should be terminated.

Block malicious application requests: Application (most commonly web) requests are intercepted for analysis by the AI engine. Those requests identified as malicious are blocked before being processed.

D. Human Interaction

Report successful and failed attacks: The control logic gathers facts from the situational awareness database regarding all knowledge it has about attacks detected, both successful and failed. A chatbot then processes the data and fills in a reporting template adding polite human conversation as necessary.

Respond to tickets: If the control logic has successfully resolved the reported problem

or it has made a promising attempt, the chatbot closes the ticket with a response to contact again if the problem persists.

8. CASE STUDY

We will now discuss a case study that shows how the different components of the ABT framework work together to defend against an attack.

During the initial hardening phase, the control logic deploys on all clients an additional trusted certificate authority (CA) and reconfigures clients to forward client traffic to a proxy, enabling both plain and encrypted traffic analysis. The control logic configures sysmon on Windows CS enabling detailed system monitoring. Finally, network traffic is analysed by the AI engine by extracting features as suggested in [33] from the situational awareness database and classifying C&C channels using data from previous LS exercises and random forest classifiers according to Känzig et al. [18].

Now, let us consider a typical case in the early stages of the exercise: the RT fools a user (UST or through GT) (i) to download a malicious payload; and (ii) to execute it on his own client (CS). As soon as the payload is executed, it (iii) contacts the C&C server and the RT now has a foothold in the Gamenet. This payload could create persistence in order to be used in later phases of the exercise to perform destructive actions by the RT.

When the user downloads and executes the payload, three events are captured in the situational awareness database: (i) from the proxy, the requests to download a malicious payload which is in this case an executable file (e.g. script, binary, library); (ii) sysmon reports the execution of an unknown payload; (iii) the payload makes contact with its C&C server, which is detected as an anomaly in the traffic analysis.

The correlation of these three events, interpreted as anomalous and suspicious, triggers the following responses at actuators by the control logic:

Malicious processes and files are identified and neutralised: the malicious process on the compromised CS is killed, the payload sent to the situational awareness analysed, inserted in the blacklist of the proxy (e.g., hash, pattern recognition), and finally deleted from the CS.

CS hardening is improved: CS software restriction policies are updated (e.g. payload hash, execution path) in order to avoid further execution of this payload on Gamenet devices.

Network rules updated: identified C&C IPs are blocked on the network firewalls and further activity to or from these IPs is considered potentially malicious.

Report generation: The control logic generates a human-readable report summarising the events related to the compromised hosts from the situational awareness database and sends it via email.

9. CONCLUSION AND OUTLOOK

We have described a novel framework that connects sensors with network, device, application and user actuators in order to automate the defence of complex cyber infrastructures against sophisticated attacks. By connecting these components together through an AI-powered computing system, we can perform automated mitigating actions to quickly and efficiently combat various attacks upon detection. Our framework is thus well suited to protect complex infrastructures which need to maintain service availability against multistage attacks. We have highlighted how our framework functions in the context of Locked Shields, but the proposed framework is applicable to other infrastructures that must maintain high service availability while under attack. We acknowledge that this paper presents only the framework and its implementation is outside the scope of the paper. Therefore, as future work, we plan to implement our architecture in order to compete in upcoming cyber defence exercises and to compare its performance with human teams. We hope that our work will also inspire other researchers in implementing similar automated cyber defence teams so that these systems could eventually compete and learn from each other.

ACKNOWLEDGEMENTS

REFERENCES

[1]   "Zeek." https://zeek.org/
[2]   "Snort." https://www.snort.org/
[3]   H. Alqahtani, I. H. Sarker, A. Kalim, S. Md. Minhaz Hossain, S. Ikhlaq, and S. Hossain, "Cyber Intrusion Detection Using Machine Learning Classification Techniques," in *Computing Science, Communication and Security*, Singapore, Jul. 2020, doi: 10.1007/978-981-15-6648-6_10.
[4]   A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," *IEEE Commun. Surv. Tutor.*, vol. 18, no. 2, 2016, doi: 10.1109/COMST.2015.2494502.

[5] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. M. Leung, "A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View," *IEEE Access*, vol. 6, 2018, doi: 10.1109/ ACCESS.2018.2805680.

[6] "Locked Shields." https://ccdcoe.org/exercises/locked-shields/

[7] K. Kasak, "Lessons learned from Locked Shields 2013 exercise," in *2nd ENISA Int. Conf. Cyber Crisis Coop. and Exercises*, Ath., Greece, 2013. [Online]. Available: https://www.enisa.europa.eu/events/2nd-enisa-conference/presentations/kaur-kasak-nato-ccdcoe-lessons-learned-from-the.pdf

[8] O. E. David, N. S. Netanyahu, and L. Wolf, "DeepChess: End-to-End Deep Neural Network for Automatic Learning in Chess," in *ICANN 2016*, 2016, doi: 10.1007/978-3-319-44781-0_11.

[9] D. Silver *et al*., "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," *Science*, vol. 362, no. 6419, 2018, doi: 10.1126/science.aar6404.

[10] E. Seker and H. H. Ozbenli, "The Concept of Cyber Defence Exercises (CDX): Planning, Execution, Evaluation," presented at Int. Conf. on Cyb. Sec. and Prot. of Dig. Serv. (Cyber Security), 2018, doi: 10.1109/CyberSecPODS.2018.8560673.

[11] E. Tyugu, "Artificial intelligence in cyber defense," presented at the 3rd Int. Conf. on Cyb. Conf., 2011, Tallinn. [Online]. Available: https://www.ccdcoe.org/uploads/2018/10/ArtificialIntelligenceInCyberDefense-Tyugu.pdf

[12] G. Bernieri, M. Conti, and F. Turrin, "Evaluation of Machine Learning Algorithms for Anomaly Detection in Industrial Networks," presented at IEEE Int. Symp. on Meas. Net., 2019, doi: 10.1109/ IWMN.2019.8805036.

[13] D. S. Bogatinov, M. Bogdanoski, and S. Angelevski, "AI-Based Cyber Defense for More Secure Cyberspace," in *Handbook of Research on Civil Society and National Security in the Era of Cyber Warfare*, IGI Global 2016, doi: 10.4018/978-1-4666-8793-6.ch011.

[14] N. Kaloudi and J. Li, "The AI-Based Cyber Threat Landscape: A Survey," *ACM Comput. Surv*., vol. 53, no. 1, 2020, doi: 10.1145/3372823.

[15] I. Chomiak-Orsa, A. Rot, and B. Blaicke, "Artificial Intelligence in Cybersecurity: The Use of AI Along the Cyber Kill Chain," in *Comput. Collect. Intel*., 2019, doi: 10.1007/978-3-030-28374-2_35.

[16] "Suricata." https://suricata-ids.org/

[17] "Anti-Virus Software." https://cs.stanford.edu/people/eroberts/cs201/projects/2000-01/viruses/anti-virus.html

[18] N. Känzig, R. Meier, L. Gambazzi, V. Lenders, and L. Vanbever, "Machine Learning-based Detection of C&C Channels with a Focus on the Locked Shields Cyber Defense Exercise," presented at CyCon 2019, doi: 10.23919/CYCON.2019.8756814.

[19] M. Ahmed, A. Naser Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *J. Netw. Comput. Appl*., vol. 60, 2016, doi: 10.1016/j.jnca.2015.11.016.

[20] P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Comput. Secur*., vol. 28, no. 1, 2009, doi: 10.1016/j.cose.2008.08.003.

[21] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, "A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection," in *SIAM Int. Conf. Data Mining,* 0 vols., SIAM, 2003.

[22] M. Landauer, F. Skopik, M. Wurzenberger, and A. Rauber, "System Log Clustering Approaches for Cyber Security Applications: A Survey," *Comput. Secur*., vol. 92, 2020, doi: 10.1016/j.cose.2020.101739.

[23] R. Alexander, "Reducing Threats by Using Bayesian Networks to Prioritize and Combine Defense in Depth Security Measures," *J. Inf. Secur*., vol. 11, no. 3, Art. no. 3, 2020, doi: 10.4236/jis.2020.113008.

[24] K. B. Lyons, "A Recommender System in the Cyber Defense Domain." MA diss., Dept. of Elect. and Comput. Eng., Grad. Sch. of Eng. and Man. Air Force Inst. of Tech., OH USA, 2014. [Online]. Available: https://scholar.afit.edu/etd/612

[25] X. Fang, M. Xu, S. Xu, and P. Zhao, "A deep learning framework for predicting cyber attacks rates," *EURASIP J. Inf. Secur*., vol. 2019, no. 1, 2019, doi: 10.1186/s13635-019-0090-6.

[26] Y. Shin and L. Williams, "An empirical model to predict security vulnerabilities using code complexity metrics," *ACM-IEEE Int. Symp. Empirical Softw. Eng. Meas*. 2008, doi: 10.1145/1414004.1414065.

[27] S. A. Abdul-Kader and J. C. Woods, "Survey on Chatbot Design Techniques in Speech Conversation Systems," *Int. J. Adv. Comput. Sci. Appl*., vol. 6, no. 7, Art. no. 7, 2015. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2015.060712

[28] E. Handoyo, M. Arfan, Y. A. A. Soetrisno, M. Somantri, A. Sofwan, and E. W. Sinuraya, "Ticketing Chatbot Service using Serverless NLP Technology," presented at the ICITACEE 2018, doi: 10.1109/ ICITACEE.2018.8576921.

[29] F. E. Office, "AI Chatbot to Realize Sophistication of Customer Contact Points," *FUJITSU Sci. Tech. J.*, vol. 54, no. 3, 2018.

[30] A. M. Rahman, A. A. Mamun, and A. Islam, "Programming challenges of chatbot: Current and future prospective," presented at IEEE R10-HTC, 2017, doi: 10.1109/R10-HTC.2017.8288910.

[31] NATO CCDCOE, "LockedShields 2013 After Action Report." https://ccdcoe.org/uploads/2018/10/LockedShields13_AAR.pdf (accessed Jan. 5, 2021).

[32] NATO CCDCOE, "LockedShields 2012 After Action Report." https://ccdcoe.org/uploads/2018/10/LockedShields12_AAR.pdf (accessed Jan. 5, 2021).

[33] "Zeek Flowmeter." https://github.com/zeek-flowmeter/zeek-flowmeter

Threat Actor Type Inference and Characterization within Cyber Threat Intelligence

Vasileios Mavroeidis
University of Oslo
Oslo, Norway
vasileim@ifi.uio.no

Ryan Hohimer
DarkLight Inc.
Richland, Washington, United States
ryan.hohimer@darklight.ai

Tim Casey
Intel Corp.
Chandler, Arizona, United States
tim.casey@intel.com

Audun Jøsang
University of Oslo
Oslo, Norway
audun.josang@mn.uio.no

Abstract: As the cyber threat landscape is constantly becoming increasingly complex and polymorphic, the more critical it becomes to understand the enemy and its modus operandi for anticipatory threat reduction. Even though the cyber security community has developed a certain maturity in describing and sharing technical indicators for informing defense components, we still struggle with non-uniform, unstructured, and ambiguous higher-level information, such as the threat actor context, thereby limiting our ability to correlate with different sources to derive more contextual, accurate, and relevant intelligence. We see the need to overcome this limitation in order to increase our ability to produce and better operationalize cyber threat intelligence. Our research demonstrates how commonly agreed-upon controlled vocabularies for characterizing threat actors and their operations can be used to enrich cyber threat intelligence and infer new information at a higher contextual level that is explicable and queryable. In particular, we present an ontological approach to automatically inferring the types of threat actors based on their personas, understanding their nature, and capturing polymorphism and changes in their behavior and characteristics over time. Such an approach not only enables interoperability by providing a structured way and means for sharing highly contextual cyber threat intelligence but also derives new information at

machine speed and minimizes cognitive biases that manual classification approaches entail.

1. INTRODUCTION

Cyber threat intelligence (CTI) is undeniably an essential element for building a robust security posture against adversarial attacks. Establishing a threat intelligence program allows security teams to benefit from increased situational awareness, and thus minimize their organizations' attack surfaces. Evidence-based knowledge of both adversary dynamics and an organization's attack surface can support anticipatory threat reduction. Organizations follow a process of increasing maturity with respect to their cyber capability, transitioning from manual and reactive approaches to more automated and proactive.

Proactive cyber defense is intelligence-driven and focuses on providing awareness and preparing an organization against anticipated attacks. Every adversarial attack can be decomposed into elements that provide information about the who, what, where, when, why, and how. The *who*, commonly known as attribution, identifies the individual, group, organization, or nation that conducted the adversarial operation. The *what* reflects the scope of the attack. The *where* relates to the attack's direction, such as where it is coming from and its target – an organization, industry, or country. The *when* can be perceived as the timestamp of the attack and can be deterministic or probabilistic. The *why* is equivalent to motivation and designates the goals and the objectives of the adversary. The *how* is made up of the tactics, techniques, and procedures (TTPs) employed by the adversary for conducting the operation. Collectively, these factors provide insight into how adversaries plan, conduct, and sustain their operations.

Attribution is typically a challenging task requiring direct evidence through principled and systematic analysis which correlates multiple internal and external data sources and threat intelligence. Such a process identifies and maps TTPs and associated tools and infrastructure to known sources of similar attacks. However, threat actors intend to remain unidentified and employ deception and obfuscation techniques that can lead to incorrect attribution or weakening the possibility of correctly associating a particular

activity with a known adversary. For example, the Russia-backed group Turla (also known as Waterbug) was discovered to be using the infrastructure and malware of APT34 (also known as OilRig), an Iranian threat group [1]. Nevertheless, many times, a threat actor profile is created and linked to one or more adversarial operations based on common identifiable properties without actual attribution, meaning that the adversary's real-world identity remains unknown.

Capturing high-level information, such as the motives behind an adversarial operation and contextualizing technical findings; for example, by estimating the sophistication level, skills, and resources needed to plan and execute the attack, can characterize the perpetrator and infer its nature even when direct attribution has not been achieved. The opposite is also plausible. The nature of a perpetrator reflects its capability, persistence, and motives. In addition, in a threat landscape that has become very diversified and hybridized, the importance of portraying adversaries and their nature as threat actor types is apparent. Threat actors are continuously evolving and are becoming polymorphic with multiple motivations and goals. Existing approaches in characterizing threat actors and their operations mostly fall under the category of regular intelligence reports that fail to capture information in a specific representation format that both humans and machines can interpret. On the other side, lies purely technical information intended to be consumed directly by cyber defense products.

A wide range of threat actor types exists, ranging from disgruntled employees to organized cyber crime and nation-state-backed groups. Threat actors have specific traits common to most of their behaviors. For example, an employee with a grudge against their organization is motivated by disgruntlement. In contrast, a state-sponsored group may aim to achieve dominance over another nation for geopolitical reasons. To operationalize this type of characterization, we need to satisfy two criteria. First, the definitions of actor types must be unambiguous, and second, we must characterize them using a set of attributes that enables robust, reliable enumeration and inference.

This research reflects the operational and strategic benefits derived from semantically portraying threat actors as threat actor types (e.g., nation-state, hacktivist, terrorist, organized cyber crime) to understand the actors' nature and capture polymorphism and changes in their behavior and characteristics over time. Furthermore, we present an ontological system for threat actor type inference which relies on a standard set of attributes for characterizing threat actors and their operations. Axioms (expressions) capture domain knowledge regarding the composition of threat actor types based on their defining attributes. The presented approach can augment existing static enumerative approaches for threat actor type classification with a flexible generative system based on the logic encapsulated in the ontologies. Such an approach enables machine understanding and logical reasoning based on that understanding with

transparent and explicable results. The proof-of-concept ontology we engineered utilizes Casey's Threat Agent Library (TAL) [2]. The original TAL typology has been refined and can be updated further to reflect a more contemporary description of threat actor types and their defining attributes.

A semantically expressed threat actor typology based on a set of standard characterization attributes provides the following advantages.

- Based on commonly agreed-upon definitions, a machine-understandable interpretation of threat actor types and their defining attributes eliminates ambiguity regarding their meaning by annotating their unique characteristics. The term *commonly* above refers to the need for interoperability. A standard vocabulary and representation for threat actor types can be integrated across different technologies, such as threat intelligence platforms and threat intelligence sharing languages, and used when generating threat reports. For example, people often interpret seemingly simple terms such as *hacktivist* differently. Correlating a threat actor type with an operation is then subject to fallacies when the semantics for what comprises a particular type are not in place. This makes shareable information inaccurate and contradictory since different entities may have different interpretations of the same term leading to inconsistent threat actor profiles.

- Representing domain knowledge in a declarative form, such as axioms and facts, can enable automatic inference via the ability of machines to reach a conclusion based on evidence. In this research, axioms capture the unique attribute combinations that characterize different threat actor types. Using a description logics reasoner, also known as an inference engine, instances of threat actors can be programmatically examined to infer their type. Automatic inference also speeds up traditional analytical processes that require competing hypotheses about the adversary's type to be tested.

- Polymorphism and changes in threat actor behavior over time are becoming common, with adversaries being influenced by different motivations and goals. Some threat actors evolve in nature and gradually engage in larger-scale and more complex operations. In contrast, others pause their operations, disappear, or even go through organizational changes like establishing new units. It is essential for the threat intelligence community to recognize and formally represent polymorphism and behavioral changes over time so that threat actor profiles can evidentially account for more than one threat actor type (Figure 1). For example, as presented in Section 5, the state-sponsored Lazarus Group has engaged in activities not only motivated by geopolitical

reasons to achieve dominance over other nations by conducting stealthy cyber-espionage campaigns but also for nationalistic reasons and revenge by engaging in destructive hacking, as well as for financially motivated reasons by conducting bank heists possibly to fund their operations. As discussed later, available threat actor knowledge bases appear to fail to capture polymorphism and behavioral changes, resulting in monolithic representations that lack evidence-based relationships concerning the derivation of their characterization. In addition, most of the time, the characterizations are based on proprietary works that are also ambiguous due to nonexistent or insufficient definitions. Ambiguity and imprecision create confusion and diminish the value of intelligence in cyber operations.

FIGURE 1: SEMANTIC MODELING OF THREAT ACTOR POLYMORPHISM
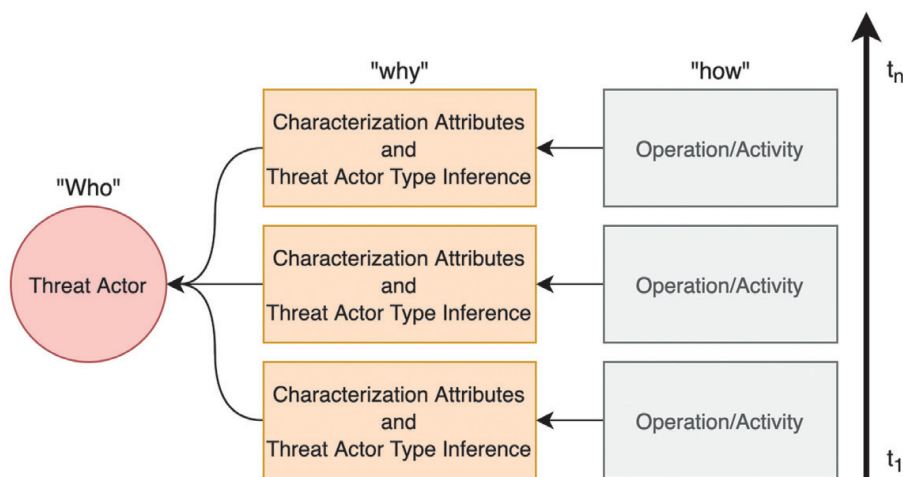
- The definition and utilization of characterization attributes (e.g., motivations, goals, objectives, visibility) can contextually enrich cyber threat intelligence and enable granular querying of higher contextual precision to answer complex questions. The derived intelligence can provide defenders with increased situational awareness and thus allow them to better prioritize their defense efforts according to their most relevant threats.

The rest of the paper is organized as follows. Section 2 introduces the Threat Agent Library [2] that was referenced to create a prototype ontology for threat actor type inference, and presents and analyzes different threat actor knowledge bases with respect to how they handle high-level contextual information in terms of ambiguity, structured shareability, explainability, and most importantly operationalization ability. Additionally, Section 2 discusses how the Structured Threat Information eXpression (STIX) language deals with interpreting threat actor polymorphism. Section 3

discusses knowledge representation and ontology engineering within the cyber threat intelligence domain, and annotates how ontology inference can provide defenders with additional information and insights at machine speed. Section 4 presents an ontology for threat actor characterization and threat actor type inference. Section 5 validates the proposed concept's efficacy and presents a use-case analysis where the ontology presented in Section 4 is used to infer threat actor types automatically. Furthermore, Section 5 demonstrates the potential of characterization attributes in providing highly contextual and queryable cyber threat intelligence. Finally, Section 6 concludes the paper.

2. BACKGROUND INFORMATION

A. Threat Agent Library

Introduced in 2007, the Threat Agent Library (TAL) [2] is a set of definitions and descriptions to represent significant threat agent categories, or as termed in this paper, threat actor types. The TAL was developed to support risk management processes by simplifying the identification of threat agent archetypes that pose the most significant risk to specific assets (Figure 2). Based on the available information on each archetype class, an organization can get an insight into current adversarial activities and consequently take action to improve its security posture. The library (Table I) enumerates twenty-one archetypes (e.g., government spy, radical activist, untrained employee, disgruntled employee) and their associated defining attributes: access, outcome, limits, resources, skills, objective, visibility, and motivation. The defining attributes reflect the typical characteristics of each threat actor type.

FIGURE 2: RISK ASSESSMENT USING THE THREAT AGENT LIBRARY
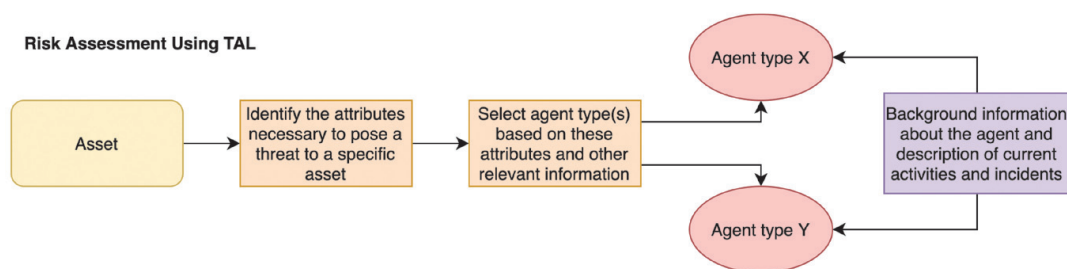
This research presents a proof-of-concept ontological representation of TAL, with minor improvements, for automatically inferring threat actor types from cyber threat intelligence instances (objects). The decision to use TAL is based on its assessment of combinations of characterization attributes that uniquely identify different threat actor types. Further, we emphasize the importance of having a set of standard characterization attributes to contextualize cyber threat intelligence, thereby making it more actionable

and relevant. We also argue that modeling approaches should be temporal-based to capture threat actor polymorphism and behavioral changes over time. As presented in the next sections, available threat actor knowledge bases struggle to capture such formalisms resulting in contextual loss and ambiguity.

B. Threat Actor Characterization Using STIX 2.1

Structured Threat Information eXpression (STIX) is a schema that defines a taxonomy for cyber threat intelligence. We discuss and analyze STIX version 2.1 [3] for two reasons. First, because of its ability to describe threat actors, threat actor activity, and their associated characteristics in a machine-readable format, and second, because it has been embraced as the standard representation format for sharing cyber threat intelligence in a structured manner.

The **STIX Threat Actor object** aggregates information about threat actors, such as their goals, motivations, sophistication, resource-level, and type. Additionally, it utilizes relationship objects to reference objects that represent the actual identity behind a threat actor (be it a human or organization), the tools that the actor has been known to use or used in a specific attack, the patterns of attack that the actor is known to follow, the location where the actor is believed to be, infrastructure both owned and compromised that the actor is known to use, as well as attributes about the actor that help characterize them. This is an object of high value in proactive cyber defense where strategic, operational, and tactical cyber threat intelligence play a significant role. Figure 3 presents the STIX threat actor object with its characterization attributes and relationships with other objects.

FIGURE 3: STIX THREAT ACTOR OBJECT
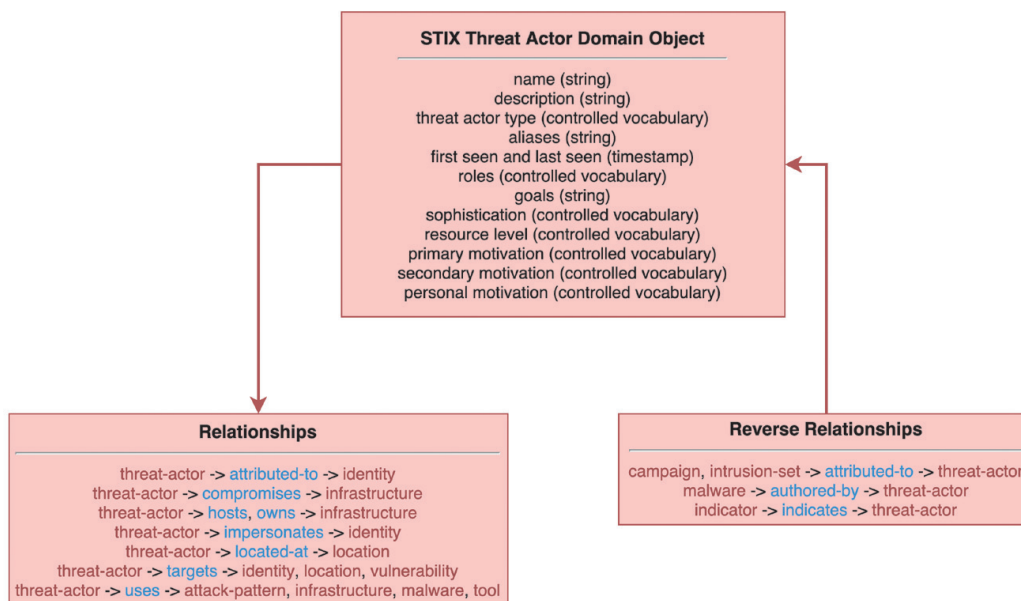
A critical aspect that the STIX threat actor object does not account for is capturing and semantically representing behavioral polymorphism in a temporal manner, as in the case where a threat actor is conducting different operations than what is known, reflecting a possible change to its primary or secondary motivations and goals. Furthermore, the characterization attributes of the threat actor object do not hold any direct relationships with other objects to justify the existing characterization. This is especially the case when a threat actor object has more than one value populated for an attribute (e.g., a threat actor that accounts for more than one threat actor type). Also, some of the STIX vocabularies used for characterizing adversaries are ambiguous because they lack definitions. The generation of the threat actor type attribute is a manual and subjective process prone to human fallacies. For example, a threat actor object with the populated threat actor type value *nation-state* and resource-level *individual* (limited resources) is unlikely to be correct but is deemed a valid STIX statement. This reflects the advantage of utilizing an automated generative threat actor type inference approach (Section 4) for augmenting existing manual approaches.

C. Threat Actor Knowledge Bases

A knowledge base is a collection of information about a particular subject area that can be used to support decision-making and draw conclusions. A knowledge base with information about threat actors' capabilities, goals and motivations, and past and ongoing activities can inform prevention and response strategies. Unstructured knowledge bases can be a simple aggregating system such as a collection of threat reports. At a basic level, the development of a structured knowledge base requires a schema that defines its structural composition, information sources for populating the knowledge base, and optimally controlled vocabularies for additional context and granular searchability. Describing a threat actor with high-confidence demands processing, correlating, analyzing, and integrating different relevant intelligence sources.

This section presents a set of open-source threat actor knowledge bases, and analyzes their structural composition with respect to how easy it is to operationalize them in the context of finding information relevant to our needs.

MITRE ATT&CK [4] is a knowledge base of known adversary tactics and techniques based on openly available analyzed activity. It is a valuable resource to better understand observed adversarial behavior, and it can be used for multiple purposes, such as for adversary emulation, behavioral analytics, cyber threat intelligence enrichment, defense gap assessment, red teaming, and SOC maturity assessment [5]. ATT&CK matrices exist about adversary behavior targeting enterprise environments, mobile, and industrial control systems. Moreover, information pertinent to the software adversaries use, mitigation techniques, procedure examples, and detection

recommendations are also available. Further, the associated PRE-ATT&CK matrix focuses on operational techniques known to be utilized before an attacker exploits a particular target network or system.

Of particular importance is the available ATT&CK Groups knowledge base, a list of known adversaries and their associated techniques and software tools. Figure 4 shows the main components of ATT&CK and their relationships.

FIGURE 4: ATT&CK MODEL RELATIONSHIPS – REDESIGNED FROM [5]
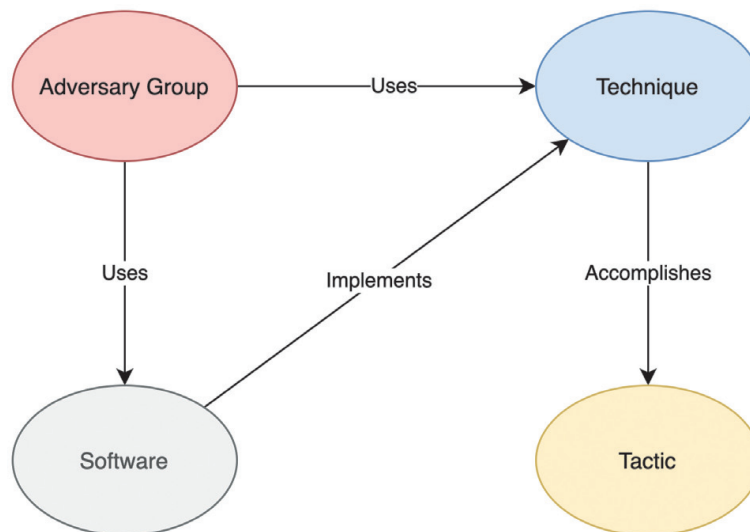
One way of getting started with ATT&CK is identifying adversarial groups relevant to an organization, based on whom they have previously targeted, such as similar organizations within the same sector, and then look at their TTPs [6]. TTPs that are commonly used can be prioritized for detection and mitigation. However, the ATT&CK Groups knowledge base lacks proper structurality and relationships between adversaries and their targets and between adversaries and their motivations. Information such as targeted countries and sectors and threat group motivations is embedded within the general description of a group and can be unstructurally searched using the ATT&CK portal. However, the vocabularies utilized to specify a group's targets and their motivations are not available, limiting searchability, and consequently, the ability to extract more relevant information. Synergistically, structuring the available information, establishing relationships between concepts, and utilizing a set of standard characterization attributes and other common vocabularies can facilitate more informed and targeted queries over the knowledge base, resulting in getting more relevant, and maybe otherwise missed TTPs to prioritize.

The description of APT19[1] is a good example of unstructured populated information regarding industries the group has targeted.

> APT19 is a Chinese-based threat group that has targeted a variety of industries, including defense, finance, energy, pharmaceutical, telecommunications, high tech, education, manufacturing, and legal services. In 2017, a phishing campaign was used to target seven law and investment firms.

Similarly, the description of APT38[2] is a good example of unstructured populated information regarding a group's motivations.

> APT38 is a financially motivated threat group that is backed by the North Korean regime. The group mainly targets banks and financial institutions and has targeted more than 16 organizations in at least 13 countries since at least 2014.

The **Threat Actor Encyclopedia** [7] is an effort from Thailand's Computer Emergency Response Team (ThaiCERT) to create a knowledge base of threat group profiles by aggregating, processing, and structuring open-source intelligence. As in other efforts, we observed ambiguity and confusion regarding the interpretation and use of characterization attributes. For instance, the threat actor encyclopedia's motivation vocabulary includes the terms *information theft and espionage*, *financial crime*, *financial gain*, and *sabotage and destruction*. Definitions of the above terms have not been provided, making it difficult, for example, to understand the contextual difference between financial gain and financial crime. It can also be argued that *information theft and espionage*, *sabotage and destruction*, and *financial crime* are not motivation types but operation types or intended effects.

The Malware Information Sharing Platform (MISP) is an open-source threat intelligence platform for collecting, storing, and sharing information about cyber security incidents [8]. Due to its open-source nature and modular architecture, the platform can integrate intelligence clusters that, in many cases, are community-driven efforts and can be used to enrich events and attributes. The **MISP Threat Actor cluster**[3] is a knowledge base of threat groups. The cluster's structural composition is an array of threat group objects that capture information related to the groups, such as name and related aliases, a description, targeted countries and sectors (e.g., private, military, government), their affiliated countries and sponsors, attribution confidence, incident types (e.g., espionage, sabotage, or defacement), references relating to the captured knowledge, relations with other groups and operations, and associated malware. A subset of the elements has been derived from the Council on Foreign

---

1   https://attack.mitre.org/groups/G0073/
2   https://attack.mitre.org/groups/G0082/
3   https://github.com/MISP/misp-galaxy/blob/main/clusters/threat-actor.json

Relations Cyber Operations[4] vocabulary used for reporting cyber incidents. Like the rest of the knowledge bases investigated, the MISP Threat Actor cluster could benefit from introducing a more expressive structured representation. Currently, multiple characterization attributes are included only in the general description of a threat actor object, making it difficult to parse the information via automated means. For instance, in the example below, the description captures information regarding the motivations, objectives, targeted countries, and the types of operations a group has been observed conducting.

> Libyan Scorpions is a malware operation in use since September 2015 and operated by a politically motivated group whose main objective is intelligence gathering, spying on influential and political figures, and operating an espionage campaign within Libya.

Moreover, the use of different non-standardized vocabularies for enriching the knowledge base and the integration of different intelligence sources for providing additional context introduces ambiguity and confusion. The two shortened examples presented below indicate the importance of utilizing a set of standard characterization attributes with accurate definitions and vocabularies for optimally resolving ambiguity and operationalizing the provided intelligence.

In the example below, *espionage* is used both to describe an incident type and a motive. Additionally, definitions for the available terms are not in place, increasing the probability of misusing the vocabularies.

```
{
   "description": "Anchor Panda is an adversary that CrowdStrike has tracked
extensively over the last year targeting both civilian and military maritime
operations...",
   "meta": {
      "attribution-confidence": "50",
      "cfr-suspected-state-sponsor": "China",
      "cfr-suspected-victims": ["United States", "..."],
      "cfr-target-category": ["Government", "..."],
      "cfr-type-of-incident": "Espionage",
      "country": "CN",
      "motive": "Espionage",
      "refs": ["..."],
      "synonyms": ["APT14"]
   },
   "value": "Anchor Panda"
}
```

---

4   https://www.cfr.org/cyber-operations/

In the example below, the motive of the group is defined as Hacktivists-Nationalists, which is reminiscent of a threat actor/group type rather than a motive that influences the actions of an actor.

```
{
  "description": "Turkish nationalist hacktivist group that has been active for
  roughly one year...The group carries out distributed denial-of-service (DDoS)
  attacks and defacements against the sites of news organizations and governments
  perceived to be critical of Turkey's policies or leadership, and purports to act in
  defense of Islam",
  "meta": {
    "attribution-confidence": "50",
    "country": "TR",
    "motive": "Hacktivists-Nationalists",
    "synonyms": ["Lion Soldiers Team", "..."]
  },
  "value": "Aslan Neferler Tim"
}
```

3. KNOWLEDGE REPRESENTATION AND ONTOLOGY

Knowledge representation conceptualizes an understanding of the world. It can provide a view of a particular domain of interest and capture that knowledge in a formal representation so that a computer system can utilize it to solve complex tasks, such as inferring new critical information. An ontology is a formalism of knowledge representation that encodes knowledge about a particular domain. An ontology is machine-understandable, holds formal semantics that carry meaning, and allows for reasoning. Formal semantics and logic ensure that the meaning of a concept is unambiguous. An ontology is defined using a knowledge representation language, such as the Web Ontology Language (OWL). An OWL ontology consists of the following three syntactic categories [9]: a sequence of logical *axioms* (statements) that are asserted to be true in the domain being described, *expressions* that represent complex notions in the domain being described (e.g., a class expression describes a set of individuals in terms of the restrictions on the individuals' characteristics), and *entities* such as classes, properties, and individuals, that constitute the basic elements of an ontology. A class represents a concept and provides the means for grouping resources with similar characteristics. For instance, a *threat actor* class can group all known adversaries. Subclasses represent concepts that are more specific than a superclass. For instance, the class *threat actor* can decompose into subclasses that

capture a threat actor's intent, such as *hostile* or *nonhostile*, and again decompose into subclasses that define hostile or nonhostile types, such as *nation-state*, *civil activist*, and *untrained employee*. Taking the Lazarus Group as an example and based on available information, it can be classified as a *nation-state* adversary, a subclass of the *hostile* class. The *hostile* class is a subclass of the *threat actor type* class, indicating that the nation-state-backed group Lazarus is an instance of a hostile threat actor. The functional syntax of this example is shown below, with Figure 5 providing an illustration.

```
Declaration ( Class( :ThreatActorType ) )
Declaration ( Class( :Hostile ) )
Declaration ( Class( :NonHostile ) )
Declaration ( Class( :NationState ) )
Declaration ( Class( :UntrainedEmployee ) )
SubClassOf ( :Hostile :ThreatActorType )
SubClassOf ( :NationState :Hostile )
SubClassOf ( :NonHostile :ThreatActorType )
SubClassOf ( :UntrainedEmployee :NonHostile )
Declaration ( NamedIndividual( :LazarusGroup ) )
ClassAssertion ( :NationState :LazarusGroup )
```

FIGURE 5: EXAMPLE ILLUSTRATION OF ONTOLOGY CLASSES AND SUBCLASSES
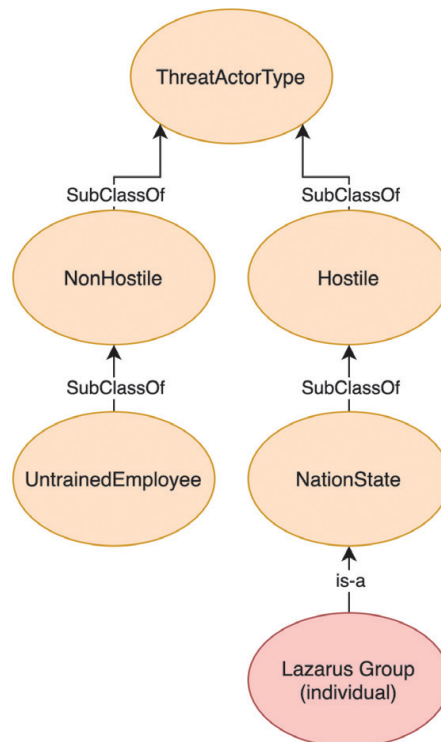
Properties define relationships between individuals (object properties) or between individuals and data type literals (data type properties). For instance, as described in the example provided in Section 2.D, APT38 is a financially motivated threat group that is backed by the North Korean regime. In addition, APT38 is also known as Stardust Chollima by Crowdstrike [10] and as BlueNoroff by Kaspersky [11]. The relation of APT38 with a particular defining motivation and other aliases can be captured by creating relevant object properties and formulating semantic triples. A triple is a set of three entities that codify a statement in the form of subject-predicate-object. This principle is illustrated in Figure 6, where the arcs represent relations (object properties – predicates), and the ellipticals represent individuals.

FIGURE 6: SEMANTIC REPRESENTATION OF APT38
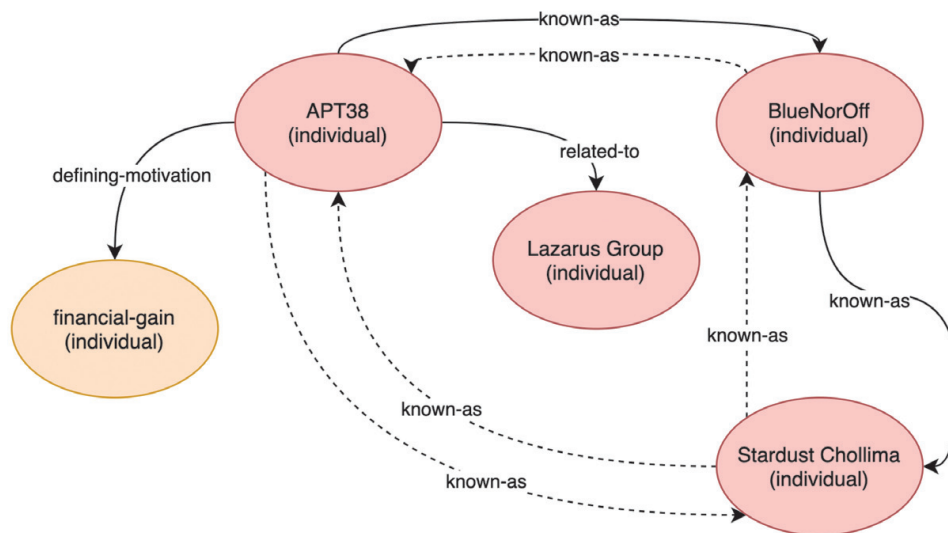
OWL offers expressive constructs for reasoning based on description logics. For example, the defined object property *known-as* is bidirectional when declared symmetric and allows traversing information when declared transitive. Property declarations can compensate for missing arcs in a knowledge base. A reasoner can parse the knowledge base and infer new information. In the example illustrated in Figure 6, the symmetric property *known-as* allows inferring that APT38 is known as BlueNoroff and the opposite, such as that BlueNoroff is known as APT38. Furthermore, because of transitivity, a reasoner infers that StarDust Chollima is also known as APT38 (dashed arc) even though it was not directly defined. Ontological axioms, expressions, and constructs can infer information based on causal relationships. For instance, a reasoner will not infer that a threat actor is of *nation-state* type when the resource-level property is not populated with the value *government*, according to the class expression that encodes what a nation-state threat actor comprises.

4. A DOMAIN ONTOLOGY FOR THREAT ACTOR PROFILING

This section presents a domain ontology for threat actor profiling and actor type inference based on the Threat Agent Library (TAL) [2]. TAL defines threat actor type attributes through controlled vocabularies, such as motivation, access, outcome, limits, resources, skills, objectives, and visibility, and when used collectively, these identify the unique characteristics of each threat actor type. Threat actor types refer to categories that adversaries can be classified into, such as spy, civil activist, and nation-state. In TAL, *threat agent* denotes a class of threat actor and is synonymous with *threat actor type*. The definitions of the TAL terms can be found in [2] and [12].

To develop the ontology, we slightly refined TAL to increase its expressiveness and resolve ambiguities that could otherwise affect ontological assertions and inferencing. TAL's threat actor types and their associated defining attributes are shown in Table I. The table's key takeaways are: TAL comprises twenty-one unique threat actor type categories and their associated characteristics based on eight attributes. The motivation attribute was added to the library in later work [12]. The shaded cells in the second column of Table I refer to either minor nonbreaking attribute modifications that resolve ambiguity concerning their ontological use, or attribute updates that allow for more flexible use. For instance, the individualistic motivation Personal Financial Gain has been replaced with Financial Gain to allow more flexible characterization, meaning that the property can now be used to characterize groups and not only individuals, such as organized cyber crime groups that operate mainly for profit, indicating financially motivated actors.

TABLE I: THREAT AGENT LIBRARY – REDESIGNED FROM [2]

Category	Attribute	Non-Hostile			Hostile																	
		Reckless Employee	Untrained Employee	Information Partner	Anarchist	Civil Activist	Competitor	Corrupt Government Official	Data Miner	Disgruntled Employee	Government Cyberwarrior	Government Spy	Internal Spy	Irrational Individual	Legal Adversary	Mobster	Radical Activist	Sensationalist	Terrorist	Thief	Vandal	Vendor
Access (1)	Internal																					
	External																					
Outcome (1-2)	Acquisition/Theft																					
	Business Advantage																					
	Damage																					
	Embarrassment																					
	Tech Advantage																					
Limits (max)	Code of Conduct																					
	Legal																					
	Extra-legal, minor																					
	Extra-legal, major																					
Resources (max)	Individual																					
	Club																					
	Contest																					
	Team																					
	Organization																					
	Government																					
Skills (max)	None																					
	Minimal																					
	Operational																					
	Adept																					
Objective (1 or more)	Copy																					
	Deny																					
	Destroy																					
	Damage																					
	Take																					
Visibility (min)	Don't Care/Any from the list opportunistically																					
	Overt																					
	Covert																					
	Clandestine																					
	Don't Care/Any from the list opportunistically																					
Defining Motivation	Accidental																					
	Coercion																					
	Disgruntlement																					
	Dominance																					
	Ideology																					
	Notoriety																					
	Organizational Gain																					
	Financial Gain																					
	Personal Satisfaction																					
	Unpredictable																					
Co-Motivation	Accidental																					
	Coercion																					
	Disgruntlement																					
	Dominance																					
	Ideology																					
	Notoriety																					
	Organizational Gain																					
	Financial Gain																					
	Personal Satisfaction																					
	Unpredictable																					

342

A high-level illustration of the ontology is presented in Figure 7. The threat actor type and characterization attribute classes enumerate possible values using individuals (instances). For example, the *visibility* attribute comprises four individuals that define different levels of visibility: clandestine, covert, opportunistic, and overt.

FIGURE 7: HIGH-LEVEL REPRESENTATION OF ONTOLOGY CLASSES AND ASSOCIATED INDIVIDUALS
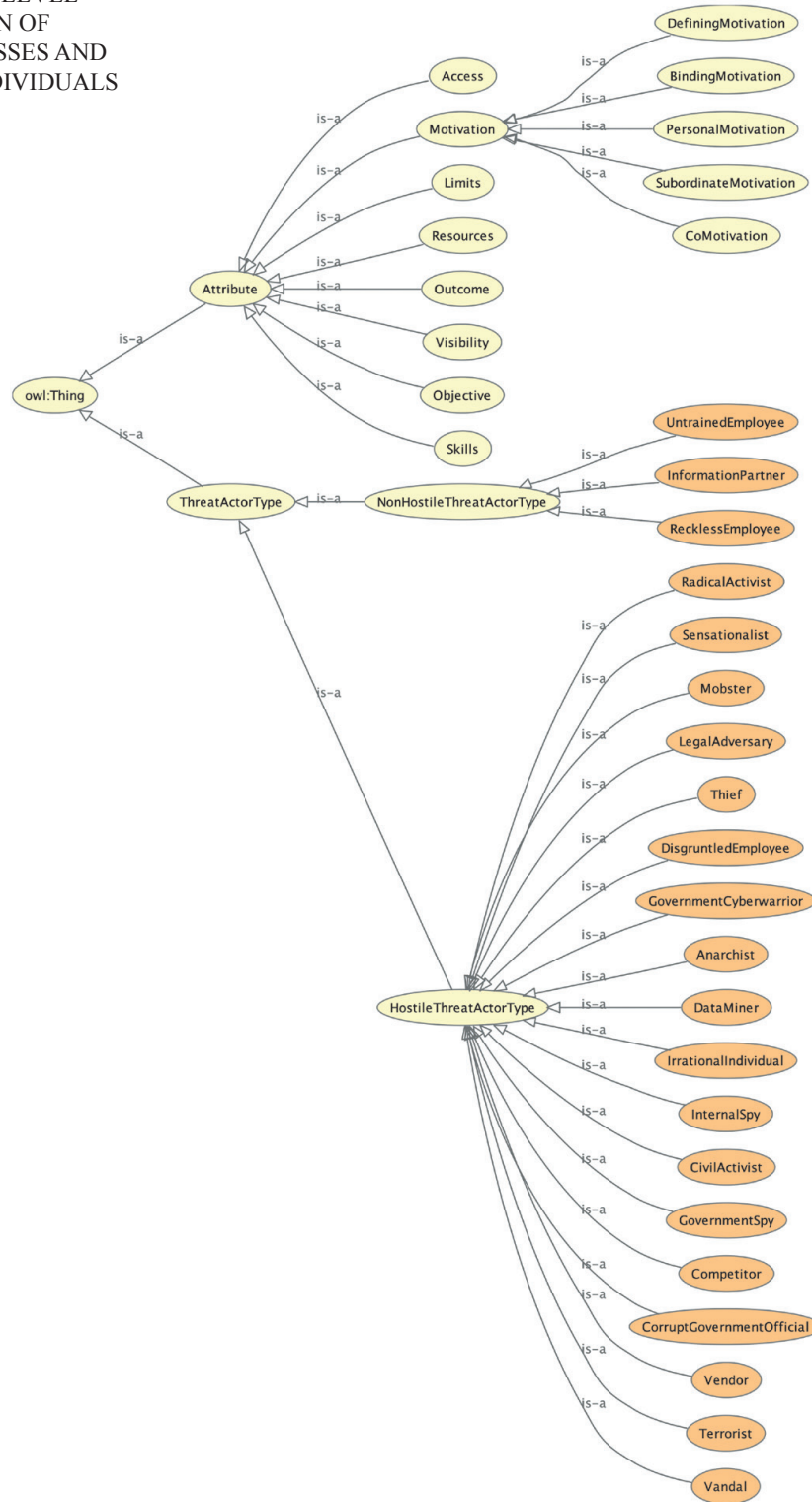
Object properties relate individuals to individuals. For example, an individual (object) that describes an adversarial operation can have a relationship to a motivation that is believed to influence the attack, such as the desire to achieve dominance. This can be expressed using the object property *hasDefiningMotivation*, deriving a semantic triple (*subject-hasDefiningMotivation-dominance*).

In addition, the ontology can automatically infer threat actor types, decreasing the human biases entailed in traditional manual classification and decision-making processes, by capturing the existing domain knowledge within ontology expressions (axioms) that characterize threat actor types based on combinations of the attributes mentioned earlier. An example expression that captures the combination of attributes comprising a nation-state-backed actor (government cyberwarrior based on TAL) is shown below in Manchester syntax.

```
((hasVisibilityAttribute some Visibility) or
(hasVisibilityAttribute value visibility:dontCare))
and ((hasObjectiveAttribute value objective:damage) or
(hasObjectiveAttribute value objective:deny) or
(hasObjectiveAttribute value objective:destroy))
and ((hasOutcomeAttribute value outcome:damage) or
(hasOutcomeAttribute value outcome:embarrassment))
and (hasAccessAttribute value access:external)
and (hasDefiningMotivationAttribute value motivation:dominance)
and (hasLimitsAttribute value limits:extraLegalMajor)
and (hasResourcesAttribute value resources:government)
and (hasSkillsAttribute value skills:adept)
```

Objects with populated attributes that fulfill expression requirements (equivalency) are classified as threat actor types in an automated manner near real-time by a description logics reasoner. As demonstrated in Section 5, polymorphic threat groups can be attributed to more than one threat actor type, compared to traditional enumerative approaches that use mutually exclusive lists and lead to contextual loss. The suggested approach does not prohibit an analyst from manually classifying a threat actor as a specific type or populating other attributes (open world assumption). Changes to the defining characterizations of threat actor types can be reflected by updating the ontology expressions. To enable temporality, the characterization attributes of a threat actor instance are populated using an individual object (instance) that connects with other related instances (e.g., malicious activity or identity) using relationships (Figure 8). Temporality-based knowledge representation can justifiably reflect shifts and polymorphism in adversarial behavior.
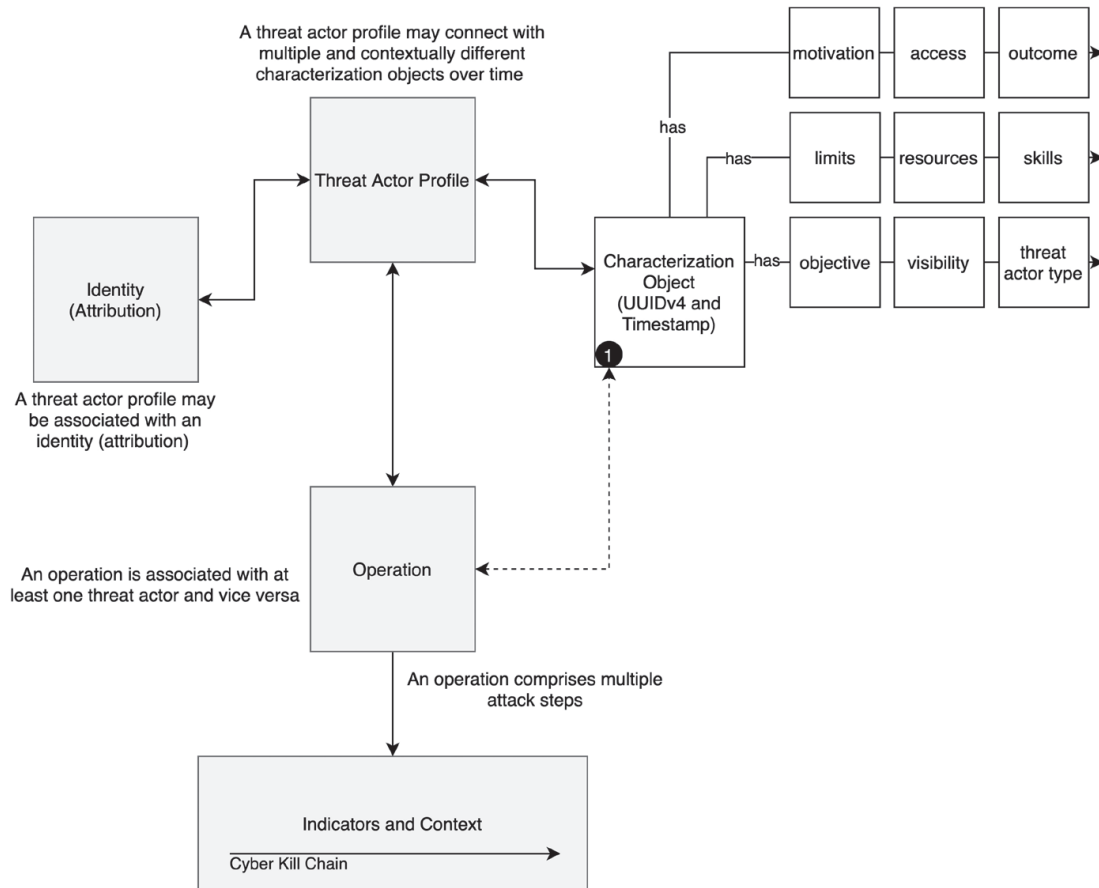
5. THE LAZARUS GROUP USE CASE

In this section, we utilize the ontology presented in Section 4 to model the Lazarus Group for the purpose of inferring threat actor types automatically. We demonstrate how a standardized set of characterization attributes for describing adversary capability and behavior makes cyber threat intelligence more contextual and queryable and makes it possible to derive new information at machine speed by utilizing a reasoner. We apply a top-down modeling approach to open-source information about operations believed to have been conducted by the Lazarus Group. Even though an attribution of high confidence has been achieved and the capabilities and sophistication of the Lazarus Group are known, we characterize the operations (use cases) based on their individual characteristics. A top-down modeling approach uses existing knowledge and historical data to create a threat actor profile and is more accurate and contextual than a bottom-up approach, which derives intelligence from early-stage ongoing analyses of cyber attacks. Nevertheless, both modeling methods should follow an evidence-

based approach by establishing direct relationships between the characterization attributes and the instances of operations the information has been derived for robust, explicable, and temporal-enabled threat intelligence.

According to the MITRE ATT&CK Groups knowledge base[5]:

> The Lazarus Group is a threat group that has been attributed to the North Korean government. North Korean groups are known to have significant overlap, and the name Lazarus Group is known to encompass a broad range of activity. Some organizations use the name Lazarus group to refer to any activity attributed to North Korea, whereas other organizations track North Korean clusters or groups such as Bluenoroff, APT37, and APT38 separately.

According to the Council on Foreign Relations[6]:

> The Lazarus Group targets and compromises entities primarily in South Korea and South Korean interests for espionage, disruption, and destruction. It has also been known to conduct cyber operations for financial gain, including targeting cryptocurrency exchanges.

The descriptions above are indicative of a polymorphic threat. Based on TAL, an ontological equivalency expression of a nation-state threat actor (government cyberwarrior) identifies the following characteristics:

- *access → external*
- *visibility → any-opportunistically*
- *objective → deny-destroy-damage*
- *limits → extra-legal, major*
- *outcome → damage, embarrassment*
- *defining motivation → dominance*
- *skills → adept*
- *resources → government*

Establishing formal threat actor type definitions using a set of machine-readable characterization attributes equips defenders with a queryable representation that can derive explicable intelligence.

The Lazarus Group is known to have been active for more than a decade and is an example of an adversary that has exhibited polymorphism and increased operational sophistication over time. The nation-state-backed group has engaged in multiple cyber espionage, destructive, disruptive, and financially motivated operations. For example,

5   https://attack.mitre.org/groups/G0032/
6   https://www.cfr.org/cyber-operations/lazarus-group

the DarkSeoul attack on March 20, 2013, targeted South Korean news agencies and banks, causing significant damage to the affected entities by wiping the hard drives of tens of thousands of computers. At an early stage, Symantec stated that the actual motives for the attacks were unclear and added that they might be part of either a clandestine attack or the work of nationalistic hacktivists taking issues into their own hands in response to political tensions on the Korean Peninsula [13]. In a report [14], McAfee, after analysis, remarked that an attack which was initially perceived as an unsophisticated incident of cyber vandalism or hacktivism had actually grown out of a sophisticated multi-year covert cyber espionage campaign that this time was indeed intended to damage, cause disruption, and potentially harvest information. Table I identifies the defining characteristics of a cyber vandal and radical activist according to TAL.

The threat actors NewRomanic Cyber Army Team and Whois Team, who claimed responsibility for the attacks in South Korea, were later discovered to be a fabrication to mask the real source of the attack. In addition, Marpaung and Lee explained that DarkSeoul was a low-tech threat compared to advanced persistent threats that nation-state groups typically perform [15].

By structuring the information about the DarkSeoul attack, the following characterization attributes emerge. The threat actor was external to the targeted entities (*access → external*) and conducted a large-scale covert operation (*visibility → covert*) which caused destruction, disruption, and possibly harvested information (*objective → destroy, damage*, and maybe *copy*). Based on the attack type and impact, we can conclude that the actor took no account of the law (*limits → extra-legal major*) and that its primary goal was large-scale data destruction with a sequential impact on the affected entities' operations (*outcome → damage*). This type of attack reflects a motivation to achieve dominance over another party, or as in this case, over another nation (*defining motivation → dominance*). Furthermore, what was initially perceived as an unsophisticated attack due to the raw destructive nature of the payload was, in fact, a coordinated strike against multiple entities delivered with precision and planning commonly associated with state-sponsored intrusion campaigns [14] (*skills → adept*), (*resources → government*). Based on the above characterization, a reasoner would infer that a government cyberwarrior conducted the operation, otherwise known as nation-state threat actor type. It is worth noting that the contextual characterization of the DarkSeoul attack in this particular case takes into account information about a set of individual attacks all described in one object, thus indicating a relatively high-level sophistication, which in turn is a factor for estimating the skills and resources required for conducting the attacks. Exemplifying each incident separately would populate objects that a reasoner would infer as the threat actor type (cyber) vandal. The attributes such as motivation, outcome, objectives, and visibility highly overlap

between the vandal and government cyberwarrior (nation-state) types. Other attributes such as skills, resources, and limits are dissimilar and annotate the differences in capability between the two types. The attribution of the DarkSeoul attack confirmed that it was planned and executed by a nation-state threat actor.

Another similar incident occurred on June 25, 2013, on the 63rd anniversary of the start of the Korean War (1950–1953), which resulted in the division of the Korean peninsula. On that day, multiple attacks reminiscent of nationalistic hacktivism, a type of patriotic activism, targeted the Blue House, government ministries, and media by defacing web pages, stealing data, and corrupting servers. One of the distributed denial-of-service (DDoS) attacks observed against the South Korean government websites was directly linked to malware used in the DarkSeoul attack [16]. The ontology in Section 4 does not account for a nationalistic hacktivist threat actor type that would ideally characterize this operation's actor. The defining attributes of each threat actor type describe their subtle differences. For example, even though the characterization attributes of the nationalistic hacktivist type would highly overlap with the radical activist type in terms of outcomes and objectives, nationalistic hacktivists are mainly motivated by the desire to achieve dominance over another nation because of their loyalty and strong devotion to their own nation or the leaders of the nation. In contrast, a radical activist operates for more ideological and political reasons to replace the fundamental principles of a society or a political system. In addition, nationalistic hacktivists would be resource-constrained compared to a nation-state-backed group. As explained in Section 3, the definition of new actor types and updating existing ones should be a standards-based task where the security community agrees on explainable characterization attribute-based descriptions for promoting and facilitating universal adoption.

In November 2014, Sony Pictures Entertainment (SPE) was attacked with malware resulting in information theft which was later used for extortion regarding canceling the release of a film depicting an assassination plot against North Korean leader Kim Jong Un. The stolen data included employee personal information, company emails, usernames and passwords, details of SPE's internal IT infrastructure, and unreleased movies. In addition, the attackers succeeded in rendering thousands of computers inoperable by deleting the master file table and the master boot record from hard drives [17]. The perpetrators identified themselves as Guardians of Peace (GOP). The attack, which was initially believed to be the work of a hacktivist group or disgruntled insiders, was later attributed to the Lazarus Group [18]. Based on available information, we characterize the operation and derive the following attributes. The Sony incident was a covert operation (*visibility → covert*) planned and executed by an unknown external group (*access → external*) that caused theft of information and damage to assets (*objective → copy, damage, destroy*). The stolen information was used to hurt

the company's image and resulted in significant financial losses (*outcome → damage, embarrassment*). The extortion demands, in addition to threatening emails sent to Sony employees, reflected a threat actor who takes no account of the law (*limits → extra-legal, major*) and an actor who attempts to achieve dominance through its actions (*defining motivation → dominance*). In addition, the threat actor demonstrated considerable resources and advanced skills, as indicated by its persistence in Sony's network and the significant losses suffered (*skills → adept*), (*resources → at least organization*). Based on the above characterization, a reasoner would infer that the populated attributes are equivalent to government cyberwarrior or otherwise known as a nation-state threat actor type. Nevertheless, the attack could also be understood as a form of nationalistic hacktivism because of its context. Interestingly, in the early stage of the attack and before the explicit demand to withdraw the movie's theatrical release, some of the targeted high-ranking Sony employees received compensation requests from the attackers for the damage they had suffered [17]. This could indicate a personal financial motivation, irrespective of the group's primary goal.

The Lazarus Group, being polymorphic, has also been observed to be financially motivated and has demonstrated highly organized and sophisticated cyber criminal behavior by penetrating targets with large financial streams. According to Kaspersky [11], Lazarus Group operations are expensive, and financially motivated attacks could be a way to better finance them. Chanlett-Avery et al. emphasized that the Lazarus Group engages in financially motivated attacks to raise revenue for the regime in response to sanctions imposed by the United States and the United Nations Security Council as a reaction to North Korea's weapons of mass destruction and ballistic missile programs, as well as human rights abuse [19].

Temporality-based semantic representation and inference provide more complete, queryable, and explainable intelligence and a certain extent of automation in intelligence generation with respect to how threat actors evolve into new behaviors. Based on the queries that an organization wants to answer, the characterization attributes and inferred information (instances) can be used to derive highly relevant and contextual cyber threat intelligence. Furthermore, universally agreed unambiguous definitions and vocabularies enable more robust information sharing.

As illustrated by Figure 9, the evidence indicates that the Lazarus Group is polymorphic and, through its operations, has exhibited behavior and capability aligned with organized cyber crime, nationalistic hacktivists, cyber vandals, and nation-state-backed entities.

FIGURE 9: THE POLYMORPHISM OF THE LAZARUS GROUP
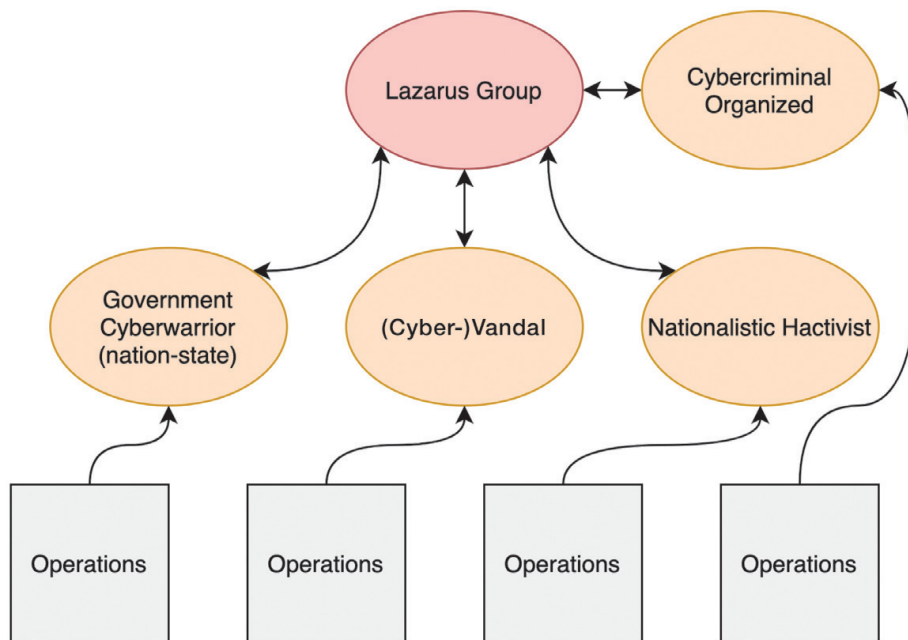
6. CONCLUSION

Threat actors are becoming increasingly sophisticated and polymorphic. To understand those hybridized threats, defenders seek timely, accurate, relevant, and actionable threat intelligence for anticipatory threat reduction. Today's threat intelligence tends to be ambiguous and inadequately structured to track and demystify changes in the behavior of actors over time, such as new goals, motivations, and related operations and TTPs. Threat actors have an asymmetric information advantage over defenders. Before executing a targeted attack, they are well aware of the profiles, infrastructures, systems, and applications of their victims. This work laid the foundation for generating highly contextual, explicable, processable, and shareable threat actor intelligence that can accurately capture, interpret, and explain changes in threat actor behavior and their polymorphism over time. In particular, we demonstrated how a set of characterization attributes can enrich threat actor information and how, in combination, can enumerate their type. By encapsulating this knowledge within an ontology, we demonstrated how a perpetrator's nature could be inferred automatically using deductive reasoning and withhold the relations/semantics that justify the inference.

ACKNOWLEDGMENTS

REFERENCES

[1] DeepSight Adversary Intelligence Team, "Waterbug: Espionage Group Rolls Out Brand-New Toolset in Attacks Against Governments," Symantec, 2019. Accessed: Jun. 2020. [Online]. Available: https://symantec-enterprise-blogs.security.com/blogs/threat-intelligence/waterbug-espionage-governments

[2] T. Casey, "Threat agent library helps identify information security risks," Intel White Paper, 2007.

[3] B. Jordan, R. Piazza, and T. Darley, Eds., OASIS Structured Threat Information Expression (STIX™) Version 2.1, OASIS – Cyber Threat Intelligence Technical Committee, 2020. [Online]. Available: https://docs.oasis-open.org/cti/stix/v2.1/cs02/stix-v2.1-cs02.html

[4] "Adversarial Tactics, Techniques & Common Knowledge (ATT&CK)," MITRE, 2020. Accessed: Jul. 2020. [Online]. Available: https://attack.mitre.org

[5] O. Alexander, M. Belisle, and J. Steele, "MITRE ATT&CK for Industrial Control Systems: Design and Philosophy," 2020. Accessed: Jul. 2020. [Online]. Available: https://collaborate.mitre.org/attackics/img_auth.php/3/37/ATT%26CK_for_ICS_-_Philosophy_Paper.pdf

[6] K. Nickels, "Getting Started with ATT&CK: Threat Intelligence," 2019. Accessed: Jul. 2020. [Online]. Available: https://medium.com/mitre-attack/getting-started-with-attack-cti-4eb205be4b2f

[7] "Threat Group Cards: A Threat Actor Encyclopedia," ThaiCERT, 2020. Accessed: Jul. 2020. [Online]. Available: https://www.thaicert.or.th/downloads/files/Threat_Group_Cards_v2.0.pdf

[8] C. Wagner, A. Dulaunoy, G. Wagener, and A. Iklody, "Misp: The design and implementation of a collaborative threat intelligence sharing platform," in *Proc. 2016 ACM on Workshop on Inf. Sharing and Collab. Secur.*, 2016.

[9] B. Motik, P. Patel-Schneider, and B. Parsia, "OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition)," W3C, 2012.

[10] A. Meyers, "Meet CrowdStrike's Adversary of the Month for April: STARDUST CHOLLIMA," Crowdstrike, 2018. Accessed: Jun. 2020. [Online]. Available: https://www.crowdstrike.com/blog/meet-crowdstrikes-adversary-of-the-month-for-april-stardust-chollima/

[11] "Lazarus Under The Hood," Kaspersky, 2017. Accessed: Jun. 2020. [Online]. Available: https://media.kasperskycontenthub.com/wp-content/uploads/sites/31/2017/04/22070418/Lazarus_Under_The_Hood_PDF_final.pdf

[12] T. Casey, "Understanding cyber threat motivations to improve defense," Intel, 2015.

[13] A. L. Johnson, "South Korean Banks and Broadcasting Organizations Suffer Major Damage from Cyberattack," 2013. Accessed: Jun. 2020. [Online]. Available: https://community.broadcom.com/symantecenterprise/communities/community-home/librarydocuments/viewdocument?DocumentKey=6859f4a7-d5c2-4a81-bbed-dfa70470e9db&CommunityKey=1ecf5f55-9545-44d6-b0f4-4e4a7f5f5e68&tab=librarydocuments

[14] R. Sherstobitoff, I. Liba, and J. Walter, "Dissecting Operation Troy: Cyberespionage in South Korea," McAfee, 2013. Accessed: Jun. 2020. [Online]. Available: https://www.mcafee.com/enterprise/en-us/assets/white-papers/wp-dissecting-operation-troy.pdf

[15] J. Marpaung and H. Lee, "Dark Seoul Cyber Attack: Could it be worse?," in *Conf. Indonesian Stud. Assoc. in Korea*, 2013.

[16] A. L. Johnson, "Four Years of DarkSeoul Cyberattacks Against South Korea Continue on Anniversary of Korean War," Broadcom, 2013. Accessed: Jun. 2020. [Online]. Available: https://community.broadcom.com/symantecenterprise/communities/community-home/librarydocuments/viewdocument?DocumentKey=edd5c93e-7160-4bf2-a15c-f1c024feb0d7&CommunityKey=1ecf5f55-9545-44d6-b0f4-4e4a7f5f5e68&tab=librarydocuments

[17] U.S. District Court, Criminal Charge, North Korean Regime-Backed Programmer Charged in Conspiracy to Conduct Multiple Cyberattacks and Intrusions, 2018. Accessed: Jun. 2020. [Online]. Available: https://www.justice.gov/usao-cdca/press-release/file/1091951/download

[18] "Operation Blockbuster: Unraveling the Long Thread of the Sony Attack," Novetta, 2016. Accessed: Jun. 2020. [Online]. Available: https://operationblockbuster.com/wp-content/uploads/2016/02/Operation-Blockbuster-Report.pdf

[19] E. Chanlett-Avery, W. L. Rosen, W. J. Rollins, and A. C. Theohary, "North Korean Cyber Capabilities: In Brief," 2017. Accessed: Jun. 2020. [Online]. Available: https://fas.org/sgp/crs/row/R44912.pdf

Self-Aware Effective Identification and Response to Viral Cyber Threats

Pietro Baroni
University of Brescia
Brescia, BS, Italy

Daniela Fogli
University of Brescia
Brescia, BS, Italy

Massimiliano Giacomin
University of Brescia
Brescia, BS, Italy

Giovanni Guida
University of Brescia
Brescia, BS, Italy

Federico Cerutti
University of Brescia
Brescia, BS, Italy
and
Cardiff University
Cardiff, Wales, United Kingdom

Francesco Gringoli
University of Brescia
Brescia, BS, Italy

Paul Sullivan
Intelpoint Inc.
Springfield, Virginia, United States

Abstract: Artificial intelligence (AI) techniques can significantly improve cyber security operations if tasks and responsibilities are effectively shared between human and machine. AI techniques excel in some situational understanding tasks; for instance, classifying intrusions. However, existing AI systems are often overconfident in their classification: this reduces the trust of human analysts. Furthermore, sophisticated intrusions span across long time periods to reduce their footprint, and each decision to respond to a (suspected) attack can have unintended side effects.

In this position paper we show how advanced AI systems handling uncertainty and encompassing expert knowledge can lessen the burden on human analysts. In detail:

(1) Effective interaction with the analyst is a key issue for the success of an intelligence support system. This involves two issues: a clear and unambiguous system-analyst communication, only possible if both share the same domain ontology and conceptual framework, and effective interaction,

allowing the analyst to query the system for justifications of the reasoning path followed and the results obtained.

(2) Uncertainty-aware machine learning and reasoning is an effective method for anomaly detection, which can provide human operators with alternative interpretations of data with an accurate assessment of their confidence. This can contribute to reducing misunderstandings and building trust.

(3) An event-processing algorithm including both a neural and a symbolic layer can help identify attacks spanning long intervals of time, that would remain undetected via a pure neural approach.

(4) Such a symbolic layer is crucial for the human operator to estimate the appropriateness of possible responses to a suspected attack by considering both the probability that an attack is actually occurring and the impact (intended and unintended) of a given response.

Keywords: *cyber threat intelligence, machine learning, artificial intelligence*

1. INTRODUCTION

The evolution of digital-enabled activities in recent years, also boosted by the COVID-19 pandemic, led to profound changes across the digital value-chain, where new challenges have emerged and have significantly affected the cyber security industry. Cyber security risks are and will become harder and harder to assess and interpret due to the growing complexity of the threat landscape, the adversarial ecosystem, and the expansion of the attack surface [1]. This will boost the spread of attacks from Advanced Persistent Threats (APTs) [2], where fleets of sophisticated attackers constantly try to gain and maintain access to networks and the confidential information that is contained within them, or to use them as a starting point for further attacks.

To illustrate the complexity of attacks from APTs, let us refer to the Lockheed-Martin *cyber kill chain* model [3], which distinguishes seven phases attackers usually follow:

1. *Reconnaissance*: Research is conducted to identify the targets appropriate to meet planned objectives.
2. *Weaponization*: Malware is coupled with an exploit into a deliverable payload.
3. *Delivery*: Malware is delivered to the target.
4. *Exploitation*: A vulnerability is exploited to gain access to the target.

5. *Installation*: A persistent backdoor is installed on the victim's system to maintain access over an extended period of time.
6. *Command and Control (C2)*: Malware establishes a channel to control and manipulate the victim's system.
7. *Actions on objectives*: After progressing through the first six phases, which might take months, the intruder, having access to the victim's system, can easily accomplish the mission goals.

The cyber kill chain model [3] also illustrates the defence options, namely: *detect, deny, disrupt* (e.g. in-line antiviruses), *degrade* (e.g. throttling the communication), *deceive* (e.g. using decoys such as honeypots), and *destroy*.

While [3] does not provide specific guidance on choosing between the various options, [4, Fig. 7.1] illustrates how defence – specifically for APTs – is an iterative process comprising three steps: *sense* (continuously sensing adversary actions), *observe* (continuously estimating intent and the capabilities of the adversary), and *manipulate* (delivering cyber deception based on observations). In this paper, we expand on the second step, the estimation of the intent and capabilities of adversaries, and embed this into the cyber threat intelligence framework (Section 2).

When focusing on cyber threat analysis, the amount of data that needs to be processed, the tempo, and the inevitable presence of adversarial actors assembles unique challenges that require advanced artificial intelligence (AI) capabilities. Our main contribution lies in Section 2, where we illustrate the desiderata for the effective usage of AI capabilities in cyber threat analysis. In the rest of the paper, we also discuss possible – albeit not all – techniques and technologies to satisfy such desiderata, most of which are based on previous work some of us directly contributed to. Our focus is on APT attacks that necessarily require a human analyst to assess the situation: less dangerous threats can be mitigated with existing tools and techniques, and this will not be part of our investigation. In Section 3 we discuss in detail the role of the human analyst,[1] who is pivotal for the success of cyber threat intelligence. Furthermore, systems need to be aware of the presence of adversarial and deceptive actors, hence in Section 4 we discuss the need for accurate quantification of uncertainty and propose a preliminary approach to this problem for raw data. In Section 5 we discuss the need to identify complex activities linked by temporal and causal relationships. Finally, in Section 6 we focus on the strategic thinking involved in choosing between alternative courses of action to manage APT attacks, in particular that which concerns unwanted side effects that might enable the attackers to acquire information of the analyst's state of knowledge and intentions, making them aware that she is aware of their attack.

---

[1] To avoid the awkwardness of strings of *he or she*, we borrow a convention from linguistics and consistently refer to a generic intelligence analyst of one sex and a generic decision-maker of the other. The female gender won the coin toss, and will represent the intelligence analyst. Attackers will always be referred to in the plural.

2. BACKGROUND AND DESIDERATA

Cyber threat intelligence is a cyclic process that analysts use to produce knowledge about weaknesses in one or more assets in an organisation that can be exploited by one or more threats. Like traditional intelligence analysis [6], [5], it comprises several steps which, merging the contributions from some of the seminal works in the field,[2] can be summarised as follows:

1. *Direction setting*: The decision-maker poses a question or requests advice (intelligence requirement): we assume this step consists in identifying APT attacks and the side effects of countermeasures.
2. *Data collection*: The analyst collects raw data – network logs from the firewalls of her organisation – into shoeboxes.
3. *Data collation*: She imposes a standard format – standardising the attributes for each log entry – to the data in the shoeboxes to create an evidence file.
4. *Data processing*: She injects useful semantics (for her task), or *schema*, in the data; for instance, by searching for classes of information such as downloads of malware.
5. *Data analysis*: She creates a case for or against the detection of APT attacks by leveraging causal links from within the data, thus building reasonable hypotheses. If under attack, she estimates the intent and capabilities of the attackers, and highlights issues with available courses of action.
6. *Dissemination*: She identifies the relevant pieces of knowledge for the decision-maker and prepares a presentation that needs to be disseminated to the decision-maker.
7. *Feedback*: She reacts to feedback from the decision-maker, who might ask for explanations or relevant details left out of the report, and that might become a new intelligence requirement.

Three main loops are identified over these steps [6]: the policy loop, which corresponds to the process leading to the identification of intelligence requirements; the foraging loop, which moves data from sources to evidence files; and the sensemaking loop, which processes data into information and knowledge shared with the decision-maker.

In this paper we focus on the first two activities associated with the sensemaking loop, namely data processing and data analysis. While dissemination and feedback are also vital for the success of the enterprise, we will not discuss these in this paper, thus silently dropping the decision-maker from the frame. An interested reader is referred to [8] and [7] for discussions on how AI can help with writing intelligence reports.

---

[2] Prunckun [5] merges dissemination and feedback, while Pace *et al.* [46] do not distinguish between information collection and data collation, and between report writing and dissemination. Prunckun [5] also names step 2 as "Information collection" following the data to wisdom hierarchy [47].

Within the scope of our study, we introduce four desiderata that AI systems need to satisfy to effectively support the analyst.

D1: Putting the analyst at the centre. While the analyst is a highly educated and skilled professional, intelligence support tools should minimise the risk of misunderstandings and allow for frequent interactions. Studies on the quality of the interaction with AI-based systems for situational understanding are scarce, especially in the cyber security domain. The analyst should be involved from the very first steps of a project and participate in all design decisions. In particular, the interface between systems and the analyst should be based on a shared ontology, familiar to the analyst but at the same time precise enough to be used by an automated system. The ontology needs to be co-designed and continuously refined on the basis of analyst feedback. Moreover, systems should allow the analyst to ask for justifications of the reasoning path followed by the systems and underlying the obtained results. The possibility to ask questions (and receive appropriate answers) contributes by providing her with the feeling of having investigated all relevant issues and checked system reasoning. In this way, the "not invented here" syndrome [9] can be avoided and trust in system advice and acceptance can significantly increase.

D2: Embracing uncertainty. There is no such thing as a perfectly certain datum in the real world: everything comes with shades of uncertainty. Traditional uncertainty estimation approaches in AI aim at quantifying it via probabilities and this can be highly misleading. Indeed, there are (at least) two different sources of uncertainty, namely aleatoric and epistemic uncertainty [10]. *Aleatoric uncertainty* refers to the variability in the outcome of an experiment which is due to inherently random effects (e.g. flipping a fair coin): no additional source of information but Laplace's daemon [11, p. 4] can reduce such variability. *Epistemic uncertainty*, instead, refers to the epistemic state of an agent; hence, it is determined by a lack of knowledge that, in principle, can be reduced on the basis of additional data.

For instance, we can create a vanilla neural network with a softmax final layer that takes as input a dataset of Portable Executable (PE) headers[4] of pieces of software labelled either *normal* or *malware*,[5] and, for any new PE header, it returns an assessment of it being *normal* or *malware*. Figure 1 illustrates the case[6] with purple dots representing normal software and blue representing malware in the considered dataset. The yellow area represents the confidence in the class prediction: the darker the yellow, the lower the confidence. The dark yellow area lies on the class boundary

---

[3] That is, a negative attitude to knowledge that originates from a source outside the own institution.

[4] Portable Executable (PE) is the format in which Microsoft Windows requires executables to be encoded. It is composed of headers and various data and code sections. For further details, https://docs.microsoft.com/en-us/windows/win32/debug/pe-format#overview (accessed 5 Dec 2020).

[5] We limit ourselves to two classes only for illustration purpose.

[6] Clearly Figure 1 does not represent a real dataset: here we simplified it substantially by generating a toy 2D dataset for clarity of presentation.

and is a manifestation of aleatoric uncertainty: pieces of software close to such a boundary have characteristics so similar that distinguishing between them is hard, and no additional data can change this. Instead, the lighter yellow areas represent regions of high confidence, and that can be the case despite the fact that no data is present there. This is a by-product of using the commonly adopted softmax approach that divides the entire space of parameters into the set of classes it is trained for (closed world), thus leaving no room for uncertainty. This is also the main characteristic exploited in adversarial machine learning (like the Fast Gradient Sign Method [12]), where a data sample can be modified imperceptibly for a human but enough for a misclassification. Consider, for example, the blue star in Figure 1, a piece of malware that was not part of our original dataset and is close to the class boundary. A very limited modification of its attributes could transform it into the purple star on its right, which would then *magically* transform it into normal software with high confidence. A clear, honest assessment of the reliability of predictions is necessary.

FIGURE 1: NORMAL SOFTWARE (PURPLE) VS MALWARE (BLUE) CLASSIFICATION WITH CONFIDENCE LEVEL USING SOFTMAX: THE BRIGHTER THE AREA (OF YELLOW), THE GREATER THE CONFIDENCE
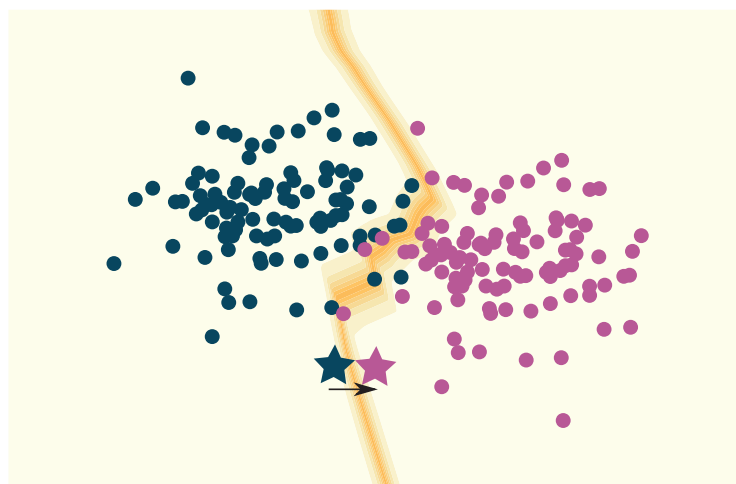
D3: Recognising complex events. An APT attack is a chain of events linked together by time and causality [3]: it is the result of a deliberate design led by human attackers. AI systems need to be equipped to reason not only about the detection of single events but also, and more importantly, to recognise events linked together by time and causality (i.e. complex events) [13]. They also need to easily adapt to evolving environments, where changes can occur in very rapid or very slow time frames. Having a perfect immutable detector of APT attacks at each stage based on past knowledge gives the analyst very little advantage in a world where new vulnerabilities are discovered each minute.

D4: Strategic thinking. When the analyst estimates the intent and capabilities of attackers, she must highlight potential side effects of the available courses of actions [3], namely: *do nothing* (always an option), *deny* (often the most common), *disrupt, degrade*, and *deceive*.[7] For the decision-maker, to choose rationally among these, the value of the information the attackers could acquire from the effects of the chosen countermeasures should be carefully pondered. It would indeed be naïve to assume that the attackers are not operating their own intelligence process. The analyst needs to consider the risk that the attackers might discover that she has some level of awareness of their operations (see the concept of *high-order theory of mind* [14]). There might also be cases where *do nothing* is a reasonable choice, like in the case of the accident at the Lawrence Berkeley Laboratory (USA) in August 1986 [15], where persistent intruders were found in a relatively low-value network as part of an "island-hopping" attack [16] towards a much higher value target. By tracing their activities for nearly one year, and employing deception warily, the attackers were found and proved to be connected to the KGB [17].

In the following sections we expand on technical solutions that can help satisfy each of the four desiderata illustrated above.

3. PUTTING THE ANALYST AT THE CENTRE

Supporting the analyst's critical thinking when facing complex, intricate menaces from APTs is not trivial. To benefit from using decision-support AI systems, the analyst must have an appropriately calibrated level of trust in the system [18], [19]. Trust is well calibrated when she sets her trust level appropriately to the AI's capabilities, accepting the output of a competent system but employing other resources or her own expertise to compensate for possible AI errors; conversely, poorly calibrated trust reduces team performance because she might trust erroneous AI outputs or not accept accurate ones [18], [20], [19].

Two problems stand out. First, it is necessary to create a stock of shared knowledge between the analyst and the artificial system she is using in order to understand the complex assessments that generally come with data analysis through machine learning and the intricate relationships between events composing an attack. Second, the analyst needs to be allowed to question the system and receive justifications for the results obtained and the reasoning processes behind them.

As far as the former issue is concerned, we argue in favour of using shared ontologies – as part of the community has already started doing; for example, [21] where domain entities and the relationships between them can be explicitly represented, thus

---

[7]   Albeit *destroy* is also an option, we will not investigate it in this paper.

clarifying the semantics for each of them. In this way, it is easier to share concepts with AI tools, as well as collecting and representing the analyst's knowledge and experience. However, querying such ontologies, thus allowing her to navigate the inevitable intricate, interrelated structures in it, soon becomes very challenging. Visual inspection is ineffective beyond a certain size threshold, while existing query languages, such as SPARQL, are often beyond the abilities of an analyst. We argue in favour of Controlled Natural Languages (CNL); in other words, "engineered subsets of natural languages whose grammar and vocabulary have been restricted in a systematic way in order to reduce both ambiguity and complexity of full natural languages" [22], while not being so restricted as a formal language. In particular, highly precise and expressive CNLs – according to the classification provided in [23] – could potentially be used as an intermediate representation between analyst and AI system, and some have also been employed in preliminary studies to facilitate human-machine joint analytical processing [24], [25], although a comprehensive assessment is still lacking. Last but not least, ontologies should also account for the uncertainty that inevitably affects all steps in the cyber threat intelligence cycle: a clear representation and communication of uncertainty plays a central role in building trust [26], [19].

As far as the latter issue is concerned, the possibility to ask the system for and obtain detailed justifications about the advice provided and the reasoning path exploited for its generation is of paramount importance. We argue that interactive interfaces must support an effective analyst-driven dialogue with the AI system, either at specific steps in the intelligence process or according to an interrupt-based protocol, where the user is allowed to ask the system at any moment during its operation. The dialogue can be based on CNL and organised according to a simple question-answer schema we co-designed [25], allowing, however, for a variety of questions that cover most of the possible information needs of the analyst, such as, for example, "justify your advice" (make the reasoning process behind the advice explicit) or "show alternatives" (illustrate possible alternative analyses and explain why they have been discarded).

4. EMBRACING UNCERTAINTY

Uncertainty pervades the entire cyber threat intelligence cycle, including the recognition of complex events linked by temporal and causal relations (Section 5) and strategic reasoning about the side effects of possible countermeasures (Section 6). Due to space constraints in this paper, we focus only on uncertainty when processing data.

The Bayesian paradigm of mathematical statistics is one of the most powerful tools we have for estimating aleatoric and epistemic uncertainty. It is based on an

interpretation of probability as a rational, conditional measure of uncertainty [27]. Pure Bayesian methods are unfeasible due to the gargantuan amount of data needed, but they can be approximated by using, for instance, Evidential Deep Learning with Noise Contrasting Estimation (EDL-NCE), which we co-designed [28], [19] and that requires less computational power and data. The main approximation behind EDL-NCE is that the posterior probability that, for instance, recalling our example from Section 2, a piece of software $\vec{x}$ is malware, $p(malware \mid \vec{x})$, is forced to be Beta-distributed or Dirichlet-distributed if we are considering more than two classes.[8] Figure 2 illustrates the result of the classification using EDL-NCE [28] on the dataset we introduced in Section 2 (see Figure 1). From a visual inspection we can appreciate how EDL-NCE derives an implicit class density estimation represented by the shades of yellow illustrating the confidence in the classification in Figure 2.

FIGURE 2: NORMAL SOFTWARE (PURPLE) VS MALWARE (BLUE) CLASSIFICATION WITH CONFIDENCE LEVEL USING EDL-NCE [28]: THE BRIGHTER THE AREA, THE GREATER THE CONFIDENCE
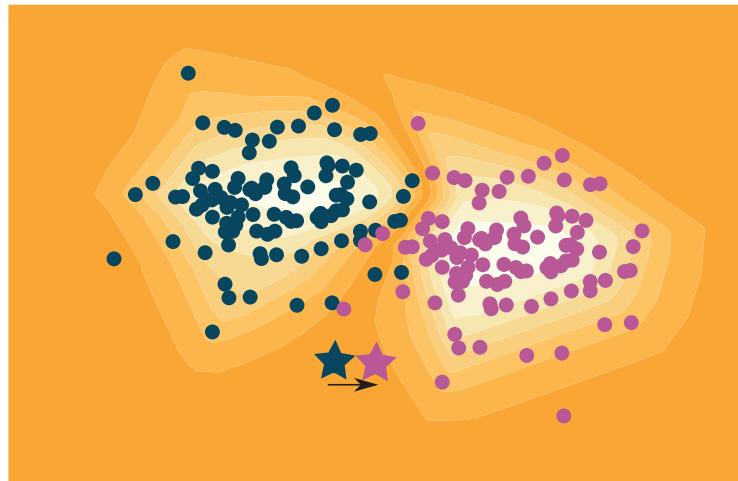

A Beta distribution needs two parameters representing the amount of evidence in favour of the two classes we are considering. For instance, let us consider again the star datapoint in Figure 2: by using EDL-NCE, we can compute $p(malware \mid \vec{x}) =$ Beta(7,5) (Figure 3(a)), which informs us that we have slightly more evidence in favour of it being a piece of malware than the opposite. With reference to Figure 3(a), we can note that (1) the expected value is 0.583, thus suggesting we are very close to the class boundary and then we have high aleatoric uncertainty, and (2) the variance

---

8   Random variables with two outcomes (e.g. tossing a coin, or detecting normal software vs malware) are known to follow the Bernulli distribution $f(\vec{y} \mid \pi) = \pi^{\Sigma y_i} (1-\pi)^{n-\Sigma y_i}$ (for $y \geq 1$). If we then want to assess the value of π from some given data samples, we can use the Bayes theorem to compute the *posterior distribution* $g(\pi \mid \vec{y}) \propto g(\pi) f(\vec{y} \mid \pi)$ on the basis of a chosen prior $g(\pi)$. Since we know that Beta distribution is the conjugate for the Bernoulli, given $g(\pi) = Beta(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1} (1-\pi)^{\beta-1}$ (for $0 \leq \pi \leq 1$), we have that $g(\pi \mid \vec{y}) \sim Beta(\alpha + \Sigma y_i, \beta + n - \Sigma y_i)$ that represents a distribution of probabilities for the phenomenon we were addressing. The generalisation of the Beta distribution to $k > 2$ outcomes (e.g. rolling a dice) is the Dirichlet distribution.

is 2.07E-2, thus suggesting a rather wide 95% confidence interval – identified by the shaded blue area under the curve of the distribution in Figure 3(a) – and then we have rather high epistemic uncertainty. Resuming the discussion presented in Section 2 (see point D2), note that a tiny manipulation of the PE headers of the star datapoint considered above might transform it into a Beta distribution with an expected value of less than 0.5; therefore, using a discriminative approach like softmax, it would be classified as normal software, but it would not have much effect on the epistemic uncertainty, and thus on the 95% confidence interval.

Beta distributions can be mapped directly into subjective logic opinions [29], namely a tuple of three values representing the *belief, disbelief*, and *uncertainty* in a given proposition. Since the three values must be non-negative, and must sum up to one, they identify a triangle in a 3D space that can easily be flattened in the 2D space depicted in Figure 3(b) as each subjective logic opinion becomes a point in the triangle [30]. In it, the vertical is the axis of confidence and it is a direct manifestation of epistemic uncertainty, from *no confidence* to *total confidence*; while the horizontal is the axis of likelihood, linked to aleatoric uncertainty, from *absolutely not likely* to *absolutely likely*. This space can be divided into different regions, each of which can be associated by a code – for example, 4C, similar to the admiralty code [5] already in use in the intelligence community – and by a couple of textual labels, such as *somewhat likely* with *some confidence*. Thanks to our previous evaluation of interfaces for decision support exposing labels representing subjective logic opinions, we argue that this has potential for creating understanding about epistemic and aleatoric uncertainty in highly-skilled personnel [25], which an analyst is supposed to be, and anecdotical evidence also suggests that decision-makers, such as physicians, appreciate the possibility of rapidly comparing the uncertainty associated with multiple reports using visual inspections of areas within the triangle. In our example, from Figure 3(b), we can see that the very same tiny manipulation that would have led a softmax approach to misclassify malware as normal software with high confidence now would not have much effect: in both cases – the original and the manipulated one – the classification shows that *chances are about even with low confidence*.

FIGURE 3: (A) GRAPHICAL REPRESENTATION OF $p(malware \mid \vec{x})$ = Beta(7,5), FOR $\vec{x}$ BEING THE STAR DATA SAMPLE IN FIGURE 2. (B) IN BLUE THE REPRESENTATION OF BETA(7, 5) AS A SUBJECTIVE LOGIC OPINION AND IN PURPLE THE REPRESENTATION OF THE CLASSIFICATION OF THE MANIPULATED INPUT INTO THE PURPLE STAR IN AN ELABORATION OF JØSANG'S [29, P. 49] SPACE OF SUBJECTIVE LOGIC OPINIONS


(a) (b)
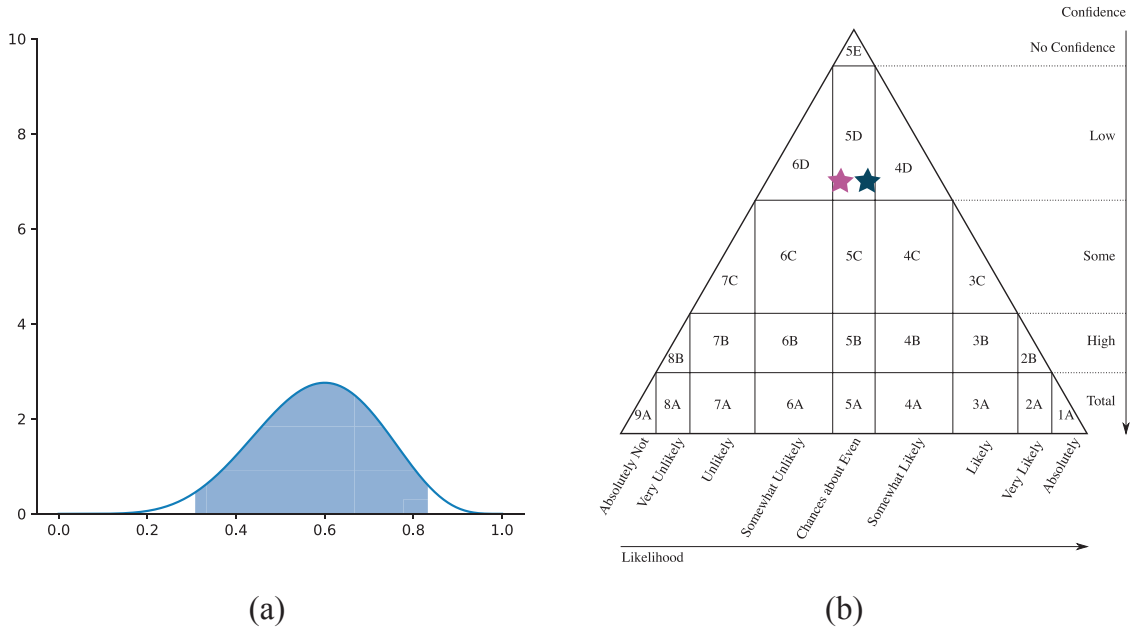
An approach to uncertainty based on subjective opinions can yield much more robust systems and, as discussed in Section 3, can help build trust with the analyst.

5. RECOGNISING COMPLEX EVENTS

Each APT attack is a set of chains of events linked together by time and causality [3] (see Section 1). We therefore advocate the use of complex event processing systems [13]. Following Luckham's [13] definitions – and differently from the everyday usage of the term event – an event is a computing object that signifies an activity that has happened. It has attributes such as the activity it represents and a timestamp or time intervals. Events can be linked by relationships of time, causality, and aggregation. If event A represents an activity that consists of the activities signified by a set of events $B_1, B_2, ..., B_n$, then A is a complex event; in other words, it is an aggregation of all the events B_i; conversely, B_i are members of A. Aggregation is a tool for making the activities in a complex system understandable to humans [13] and is the fundamental component in an *event abstraction hierarchy* that induces a sequence of levels such that the events in each level are defined on the basis of an aggregation of events at previous levels via aggregation rules. Clearly this applies to all the levels except the first one – conventionally Level 0 – which does not contain complex events.

For instance, the analyst might start looking into the evidence file containing network traces of HTTP(S) connections. Figure 4 illustrates the situation we specifically devised for this research: on the left there are the network logs collected in the evidence file. Activities are transformed into events thanks to a neural network trained to detect the type of URL, which might be a NEWS outlet, or a SEARCH engine, or the beginning of a download – an EXEC file, or even a PHISHING website (i.e. cloning a website to pose as it while delivering a malicious payload). Events might trigger the creation of other events: in the figure, downloading an EXEC has led to analysing it using a malware detector similar to the one we illustrated in Section 4, and that concluded that it is *likely with high confidence* (`certainty: 3B`) that it is a trojan; that is, malware misleading the user about its intent. An aggregation rule is then triggered and generates a complex event that signifies the detection of the delivery of a weapon as part of possible APT attacks.

FIGURE 4: ILLUSTRATION OF DETECTING THE DELIVERY OF AN APT WEAPON AS A COMPLEX EVENT. TIME FLOWS FROM TOP TO BOTTOM. ARROWS MARKED WITH ▨ REPRESENT DATA PROCESSING VIA NEURAL NETWORKS. ATTRIBUTES ARE REPRESENTED IN JSON-LIKE SYNTAX. ARROWS LABELLED WITH *CAUSES* REPRESENT CAUSAL RELATIONSHIPS BETWEEN EVENTS. BLACK DIAMONDS REPRESENT AGGREGATION RULES.
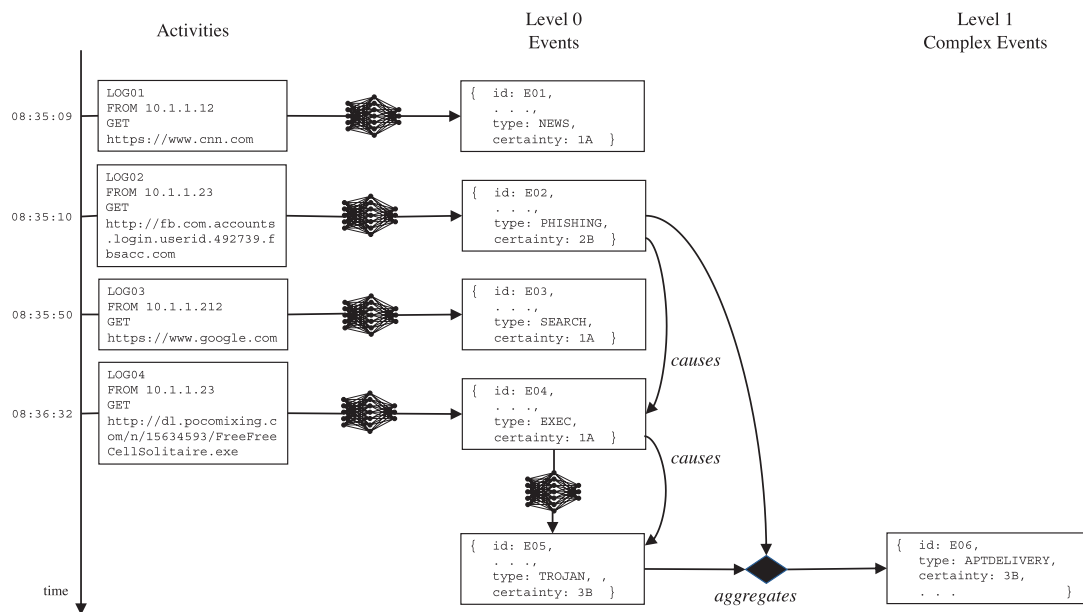
Two limiting factors emerge. The first is the identification of relationships between events, in particular the aggregation rules. They can either be elicited by domain experts, or they can be learnt from annotated datasets of sequences of events by leveraging, for instance, inductive logic programming [31]. However, it is beyond doubt that high-quality [32] domain knowledge expressed as aggregation rules must

be curated and maintained, and this adds additional weight to the usage of suitable (controlled natural) languages (see Section 3).

The second problem is linked to adaptability to new weapons. Existing approaches using complex event processing for detecting APT attacks (e.g. [34] and [33]) assume that each of the models used to create events is separately trained on top of existing curated datasets: in the world of viral threats, this lack of adaptability is unsustainable. Adaptability to new contexts is the basis of novel, neuro-symbolic approaches to (simplified) complex event processing [35], [36]. Such approaches, which we co-designed, require as input only raw pieces of information (the logs identifying the activities, left of Figure 4), sets of aggregation rules, and the final labels. By leveraging approaches such as [37], linking together symbolic knowledge (aggregation rules) with sub-symbolic data processing (the identification of events from raw data), they can train classifiers for the various events of interest, such as PHISHING, EXEC and TROJAN, without the need to provide specific information about them. Using synthetically generated raw data, we gathered evidence in favour of this, although much more is left to do.

The identification of an event abstraction hierarchy is paramount for integrating not only SIGINT,[9] but also unstructured or semi-structured OSINT,[10] for instance, by fusing activity reports from both Clearnet and Darknet. By focusing on causal and aggregation relationships, this climbing of the semantic ladder is argumentative and adheres to the best practices of critical thinking [38], [39].

6. STRATEGIC THINKING

Let us now assume that a threat has been detected. The analyst now needs to estimate the intent and capabilities of attackers, and highlight issues with possible courses of action [3]. Strategic thinking is thus needed in choosing between alternative courses of action to manage APT attacks, in particular in respect to unwanted side effects that might enable the attackers to acquire information on the analyst's state of knowledge and intentions, making them aware that she is aware of their attack. For simplicity, let us consider only *deny* and *deceive*.

We therefore argue that it is necessary to explicitly acknowledge the existence of a communication link with the attackers, who will receive information about (1) our analyst's defence capabilities for detection and (2) the value of the resources she is protecting. We can thus leverage AI techniques such as Controlled Query Evaluation

---

9   SIGINT—SIGnal INTelligence—includes either individually or in combination all communications intelligence, electronic intelligence, and instrumentation signals intelligence, in whatever way transmitted.

10   OSINT—Open Source INTelligence—includes media, internet (both Clearnet and Darknet), governmental data, professional publications as well as grey literature such as technical reports or preprints, commercial data, etc.

(CQE) [40], where an agent is defending some knowledge encoded in a database against an attacker who can perform queries on it. The defender can choose how to answer such queries but is compelled to obey secrecy constraints.

For instance, let us assume the analyst uses the graph database illustrated in Figure 5, where boxed labels represent classes, while unboxed ones represent objects instantiating *(is-a)* a class. We assume the existence of the *causedBy*, *composedBy*, and *from* relationships among objects; their semantics is linked to the complex event processing procedure illustrated in Section 5; for instance, E04 is *an* EVENT derived *from* LOG04, which *is an* ACTIVITY. Let us also assume that our ability to detect that the downloaded file is a trojan (event E05) is based on a machine learning algorithm trained on a dataset containing raw data about possible malware that we carefully created within our organisation, and that it is in the company's interest to keep it private. This is represented in Figure 5 by a black box (PRIVATE DATASET) since it contains *black knowledge*; that is, knowledge that should not be disclosed.

Terminating the attack by shutting the connection (*denying*) after E05 and triggering the creation of E06, which makes the analyst aware of the presence of an attack, might seem a reasonable choice. This, however, might signal the attackers that the analyst has access to superior knowledge – compared to the community – about the used malware. She needs to assume that the attackers are also able to detect E02 and E04, as the target interacted with a remote server at least partially under their control. By using, for example, CQE, over a probabilistic version of the graph database illustrated in Figure 5, we can now answer the question: what is the probability that in revealing E06 we also reveal E05? Since E06 builds on E02, which is informed by a public dataset, there is a reasonable argument suggesting that denying the attack can be explained only on the basis of publicly available information. For instance, all downloads from such URLs can be quarantined or sandboxed.

FIGURE 5: GRAPH DATABASE REPRESENTING THE DETECTION OF THE DELIVERY OF AN APT WEAPON (SEE FIGURE 4)
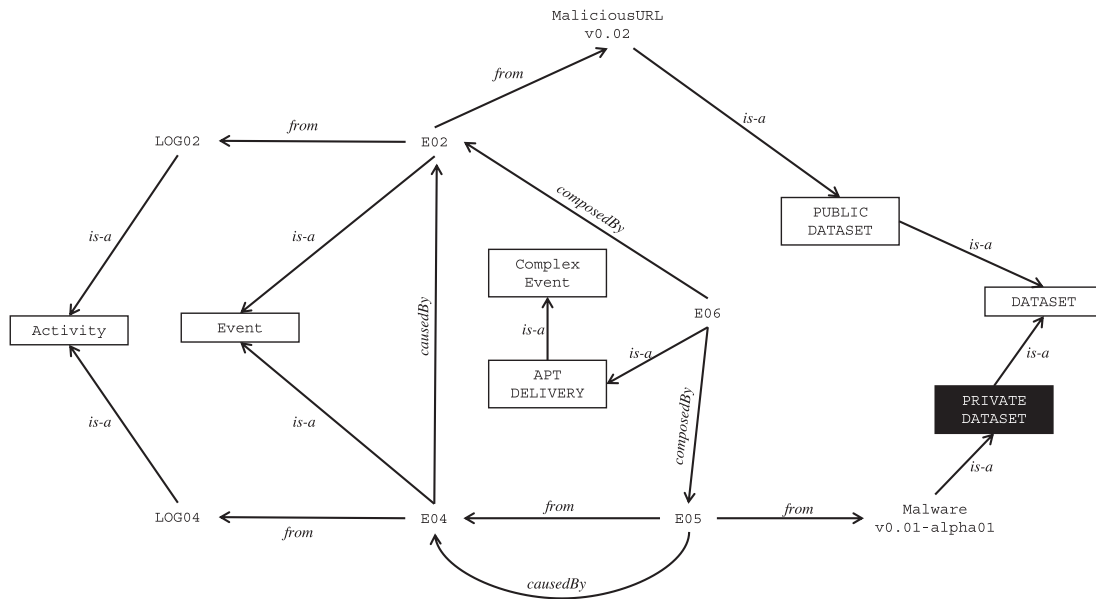
Inheriting uncertainty quantification about events and complex events from the previous processing and analysis steps (see Sections 4 and 5) and performing CQE over probabilistic knowledge bases makes it possible to encompass both epistemic and aleatoric uncertainties, as we showed in [41] and [42], thus providing the analyst with a computational mechanism for risk evaluation. This can also be used to derive a utility function to be used in state-of-the-art game theory approaches for choosing mitigation techniques [44], [43, Chs. 4, 5], [4]. These topics are, however, beyond the limits of this paper.

To conclude, considering at least one level into the high-order theory of mind coupled with epistemic and aleatoric uncertainty brings us closer to the real world, while at the same time revealing the complexity of the task and the need for self-aware artificial intelligence.

7. CONCLUSIONS

The threat landscape, adversarial ecosystem, and expansion of the attack surface all together link to an environment of staggering complexity where viral threats affect the entire fabric of our interconnected world. Optimising for the known threats only is not enough: we need to build resilient systems that embrace uncertainty and adapt to new types of complex attacks.

In this paper we embed defence against APT attacks into the cyber threat intelligence framework, and illustrate how self-aware AI tools can be used for building resilience and lessening the burden on the human analyst, who must always be at the centre of the design process.

We show that uncertainty-aware learning and reasoning can be an effective method for anomaly detection, which can provide human operators with alternative interpretations of data and accurate assessments of their confidence. This reduces misunderstandings and builds trust, while also reducing attackers' options for camouflage. Event processing algorithms can identify attacks spanning long intervals of time, which would remain undetected even by state-of-the-art intrusion detection systems. Finally, climbing the ladder of semantics is crucial for estimating the appropriateness of different responses to a suspected attack, and the impact (intended and unintended) of a given response.

Several avenues are ahead of us, including further experimental analyses that are already planned, but here we would like to mention one in particular that, due to space constraints, we left out, but that must be remarked upon. Although in this paper we implicitly assumed a centralised approach – that is, an analyst or a group of analysts overseeing the cyber infrastructure of a large organisation – the reality is that the staggering complexity of the task might require a more distributed approach, which can be achieved by as much as possible empowering autonomous agents at the edge of the network to collaborate with a single intent: a sort of *team of teams* [45] with the same purpose and shared knowledge. To this end, a possible strategy is to couple each analyst with an autonomous surrogate that, via reinforcement learning, could approximate the decision of its human counterpart and thus reduce even further the burden on human experts especially for the most trivial tasks.

ACKNOWLEDGEMENTS

Defence or the U.K. Government. The U.S. and U.K. Governments are authorised to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] ENISA, "ENISA Threat Landscape 2020 – Emerging Trends," 2020.

[2] P. Chen, L. Desmet, and C. Huygens, "A study on advanced persistent threats," in *IFIP Int. Conf. Commun. Mutlimedia Secur.,* 2014, pp. 63–72.

[3] E. M. Hutchins, M. J. Cloppert, R. M. Amin et al., "Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains," in *Proc. 6th Int. Conf. Inf. Warfare and Security*, 2011, pp. 113–125.

[4] C. A. Kamhoua, L. L. Njilla, A. Kott, and S. Shetty, Eds., *Modeling and Design of Secure Internet of Things*. USA: Wiley, 2020.

[5] H. Prunckun, *Scientific Methods of Inquiry for Intelligence Analysis*. Lanham, MD, USA: Rowman & Littlefield, 2015.

[6] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proc. Int. Conf. Intell. Anal.*, vol. 5, 2005, pp. 2–4.

[7] A. Toniolo et al., "Supporting reasoning with different types of evidence in intelligence analysis," in *Proc. AAMAS 2015*, 2015, pp. 781–789.

[8] F. Cerutti, T. J. Norman, A. Toniolo, and S. E. Middleton, "CISpaces.org: From fact extraction to report generation," in *Proc. COMMA 2018*, 2018, pp. 269–280.

[9] D. Antons and F. T. Piller, "Opening the black box of 'Not Invented Here': Attitudes, decision biases, and behavioral consequences," *Acad. Manag. Perspect.*, vol. 29, no. 2, pp. 193–217, 2015.

[10] S. C. Hora, "Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management," *Reliab. Eng. Syst. Saf.*, vol. 54, no. 2, pp. 217–223, 1996.

[11] P. S. Laplace, *A Philosophical Essay on Probabilities*. New York, NY, USA: John Wiley & Sons, 1902.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. ICLR2015*, 2015.

[13] D. C. Luckham, *The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems*. USA: Addison-Wesley Longman Publishing Co., Inc., 2002.

[14] H. Wimmer and J. Perner, "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception," *Cognition*, vol. 13, no. 1, pp. 103–128, Jan. 1983.

[15] C. Stoll, "Stalking the Wily Hacker," *Commun. ACM*, vol. 31, no. 5, pp. 484–497, 1988.

[16] S. Jajodia, P. Liu, V. Swarup, and C. Wang, Eds., *Cyber Situational Awareness: Issues and Research*. New York, NY, USA: Springer, 2010.

[17] C. Stoll, *The Cuckoo's Egg: Tracking a Spy through the Maze of Computer Espionage*. New York, NY, USA: Pocket Books, 1989.

[18] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1. Human Factors and Ergonomics Society, pp. 50–80, 2004.

[19] R. Tomsett *et al.*, "Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI," *Patterns*, vol. 1, no. 4, p. 100049, Jul. 2020.

[20] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *Int. J. Man. Mach. Stud.*, vol. 27, no. 5–6, pp. 527–539, Nov. 1987.

[21] V. Mavroeidis and S. Bromander, "Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence," *Proc. EISIC 2017*, pp. 91–98, 2017.

[22] R. Schwitter, "Controlled Natural Languages for Knowledge Representation," in *Proc. Coling 2010: Posters*, pp. 1113–1121, 2010.

[23] T. Kuhn, "A Survey and Classification of Controlled Natural Languages," *Comput. Linguist.*, vol. 40, no. 1, pp. 121–170, 2014.

[24] D. Braines, D. Mott, S. Laws, G. de Mel, and T. Pham, "Controlled English to facilitate human/machine analytical processing," in *Next-Generation Analyst*, vol. 8758, no. 7, pp. 875–808, 2013.

[25] D. Braines *et al.*, "Subjective Bayesian Networks and Human-in-the-Loop Situational Understanding," in *GKR 2017: Graph Structures for Knowledge Representation and Reasoning*, 2018, pp. 29–53.

[26] G. Bansal, B. Nushi, E. Kamar, W. Lasecki, D. Weld, and E. Horvitz, "Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance," in *Proc. HCOMP2019*, 2019, pp. 2–11.

[27] J. M. Bernardo, "Bayesian statistics," in *International Encyclopedia of Statistical Science*, Courcier, Ed., Springer, 2011, pp. 107–133.

[28] M. Sensoy, L. Kaplan, F. Cerutti, and M. Saleki, "Uncertainty-Aware Deep Classifiers using Generative Models," in *Proc. AAAI2020*, 2020, pp. 5620–5627.

[29] A. Jøsang, *Subjective logic: a formalism for reasoning under uncertainty*. Switzerland: Springer, 2016.

[30] F. Cerutti, L. M. Kaplan, T. J. Norman, N. Oren, and A. Toniolo, "Subjective logic operators in trust assessment: an empirical study," *Inf. Syst. Front.*, vol. 17, no. 4, 2015.

[31] S. Muggleton, "Inductive logic programming," *New Gener. Comput.*, vol. 8, no. 4, pp. 295–318, 1991.

[32] G. Guida and G. Mauri, "Evaluating performance and quality of knowledge-based systems: foundation and methodology," *IEEE Trans. Knowl. Data Eng.*, vol. 5, no. 2, pp. 204–224, 1993.

[33] X. Jin, B. Cui, J. Yang, and Z. Cheng, "An adaptive analysis framework for correlating cyber-security-related data," in *Proc. AINA 2018*, 2018, pp. 915–919.

[34] A. Benzekri, R. Laborde, A. Oglaza, D. Rammal, and F. Barrère, "Dynamic security management driven by situations: An exploratory analysis of logs for the identification of security situations," in *Proc. CSNet 2019*, 2019, pp. 66–72.

[35] M. R. Vilamala *et al*., "A hybrid neuro-symbolic approach for complex event processing (extended abstract)," 2020.

[36] T. Xing *et al*., "Neuroplex: learning to detect complex events in sensor networks through knowledge injection," in *Proc. SenSys2020*, 2020, pp. 489–502.

[37] R. Manhaeve, S. Dumančic, A. Kimmig, T. Demeester, and L. De Raedt, "DeepProbLog: Neural Probabilistic Logic Programming," in *Proc. NIPS2018*, 2018, pp. 3749–3759.

[38] N. Hendrickson, "Critical thinking in intelligence analysis," *Int. J. Intell. CounterIntell.,* vol. 21, no. 4, pp. 679–693, 2008.

[39] D. Walton, *Fundamentals of Critical Argumentation*. Cambridge, UK: Cambridge University Press, 2005.

[40] G. L. Sicherman, W. De Jonge, and R. P. Van de Riet, "Answering queries without revealing secrets," *ACM Trans. Database Syst.*, vol. 8, no. 1, pp. 41–59, 1983.

[41] F. Cerutti *et al*., "Obfuscation of semantic data: Restricting the spread of sensitive information," in *Proc. DL2014*, 2014, pp. 434–446.

[42] F. Cerutti *et al*., "When is Lying the Right Choice?," in *Proc. 1st Workshop on Lang. and Ontologies*, London, UK: Association for Computational Linguistics, 2015.

[43] E. Al-Shaer, J. Wei, K. W. Hamlen, and C. Wang, Eds., *Autonomous Cyber Deception*. Switzerland: Springer International Publishing, 2019.

[44] J. Pawlick, E. Colbert, and Q. Zhu, "A Game-Theoretic Taxonomy and Survey of Defensive Deception for Cybersecurity and Privacy," *ACM Comput. Surv.*, vol. 52, no. 4, Aug. 2019.

[45] S. A. McChrystal, T. Collins, D. Silverman, and C. Fussell, *Team of Teams: New Rules of Engagement for a Complex World*. New York, USA: Portfolio/Penguin, 2015.

[46] C. Pace *et al*., *The Threat Intelligence Handbook: A Practical Guide for Security Teams to Unlocking the Power of Intelligence*. CyberEdge, 2018.

[47] R. L. Ackoff, "From data to wisdom," *J. Appl. Syst. Anal.*, vol. 16, no. 1, pp. 3–9, 1989.

Quantum Communication for Post-Pandemic Cybersecurity

Martin C. Libicki
Distinguished Visiting Professor
Center for Cyber Security Studies
U.S. Naval Academy
Annapolis, MD, United States
libicki@usna.edu

David Gompert
Special Advisor
Ultratech Capital Partners
United States
Davidgompert@yahoo.com

Abstract: Current approaches to cybersecurity will become increasingly inadequate as the use of networks grows and hacking becomes more skilled. One response to this problem lies in quantum technologies. In particular, the extreme sensitivity of quantum communication makes interference readily detectable and can provide secure encryption-key distribution. However, this is likely to benefit primarily high-value networks that use encryption, leaving insecure the growing use of mass networks for distributed work. The options to approaching this conundrum are (1) to accept where quantum technology leads, (2) to accelerate the technology in general without regard to how it is used, or (3) to push the technology to include mass use. We recommend a public-private strategy for the United States and its allies to effect both high-end and mass use.[1]

Keywords: *quantum communication, technology policy*

1. INTRODUCTION

As use of the Internet and other data networks has grown, reliable, economic, and durable cybersecurity has proven elusive. While huge funding is shoveled into cybersecurity, the incidence, severity, and costs of cybercrime are escalating. The insecurity is especially acute for people, societies, and nations that rely on free politics, markets, and speech—that is, the United States and its allies—under threat from two main adversaries, Russia and China. Recent disclosures of Russian intrusion

---

[1] The authors would especially like to acknowledge Professor Nathalie de Leon of the Princeton Quantum Initiative, Princeton University, whose input on quantum science has been invaluable. Errors in this paper are, of course, ours.

into important U.S. government systems reveal that cybersecurity, despite expanding investment, has not kept pace with increasingly sophisticated hacking.

Meanwhile, the growth in network use in order to enable distributed work will persist after the pandemic. Because cybercrime increases at about the same rate as network use, and home computers are becoming substitutes for more secure workplace ones, cybersecurity will become even more challenging and expensive, yet ever more vital.

One way to help close the yawning gap between cyber vulnerability and security may lie in quantum technologies. Although practical quantum *computing* is at least a decade away, the dawn of quantum *communication* is here. The sensitivity of quantum transmissions allows hostile interference to be revealed and thereby ensures the safe passage of messages, notably those involved in encryption-key distribution. This raises the prospect of a hack-resistant "Quantum Internet," initially instantiated as secure quantum links within today's digital Internet. A Quantum Internet would not require replacing most of the Internet's infrastructure, and the cost would mostly be borne by those willing to pay for genuine cybersecurity, albeit with a focused government role.

Below, we explain the need for a public-private strategy involving U.S.-allied collaboration to guide investment, overcome technical hurdles, secure high-value networks, and extend the benefits of quantum communication to general public use.

A. Taking Stock

In a nutshell, the use of networks is accelerating; the volume and sophistication of hacking is increasing at the same rate, if not faster; return on investment in cybersecurity is generally discouraging; and the damage can be expected to grow, especially as Russia and China become more aggressive.

Remote work, prompted by the pandemic, has been both efficient and popular with employees and employers alike. If, say, half the growth in remote work due to the pandemic were to remain after the pandemic ends, network use could be up about around 25% from pre-pandemic levels (over and above baseline growth). Adding to the shift of jobs from office to home is the replacement of on-site meetings with off-site ones.

This trend is occurring not only in everyday networks but also in sensitive ones. Valuable intellectual property, such as chip designs, drug formulas, and patent applications, may be exchanged online. A great deal of unclassified but critical government business will be done remotely. The National Security Agency (NSA)

warns that the dispersal of U.S. government work to home offices presents "countless opportunities" for hacking, especially by Russian agents.[2]

Cyberattacks are escalating in proportion to network use.[3] The FBI reports that complaints about cybercrime have increased by 300% during the pandemic, and one cannot be sanguine about post-pandemic cybersecurity.[4] Investment in cybersecurity has been rising fast, from $3 billion in 2004 to $124 billion[5] in 2019, and shows no signs of slowing down.[6] Yet worldwide costs of cybercrime and cyberconflict have been rising at an even faster rate than Internet use has; by one estimate, $600 billion a year is being lost.[7] Though there are always particular successes, the macroeconomics of cybersecurity are generally unpromising.

Whether in time spent, lines of code written, people employed, or funds expended, the effort and expense required to protect, detect, patch, work around, and recover from attack far exceed those of hacking. At its higher levels, investment in cybersecurity shows sharply diminished returns.[8] Firms typically experience a flattening of the curve that relates cybersecurity achieved to cybersecurity investment.[9]

This has been so because the Internet was designed as an open utility to afford access to information, facilitate sharing, and enable collaboration. Open networks tend to have increasing as opposed to decreasing returns on investment, as adding participants benefits those already participating—an economic phenomenon favoring open networks that has propelled the digital revolution.[10] Yet protecting user-friendly systems tends to be harder than invading them, all else being equal. Conversely, the more restrictive networks are for the sake of security—access control lists come to mind—the less useful they may be for users.

2   Lily Hay Newman, "The NSA Warns That Russia Is Attacking Remote Work Platforms," *Wired*, December 7, 2020, https://www.wired.com/story/nsa-warns-russia-attacking-vmware-remote-work-platforms/.

3   Accenture, *Ninth Annual Cost of Cybercrime Study*, 2019, https://www.accenture.com/_acnmedia/pdf-96/accenture-2019-cost-of-cybercrime-study-final.pdf.

4   Catalin Cimpanu, "FBI Says Cybercrime Reports Quadrupled during COVID-19 Pandemic," ZD Net, April 18, 2020, https://www.zdnet.com/article/fbi-says-cybercrime-reports-quadrupled-during-covid-19-pandemic/.

5   Gartner, "Gartner Forecasts Worldwide Security and Risk Management Spending Growth to Slow but Remain Positive in 2020," Gartner Newsroom, June 17, 2020, https://www.gartner.com/en/newsroom/press-releases/2020-06-17-gartner-forecasts-worldwide-security-and-risk-managem.

6   "We anticipate 12–15 percent year-over-year cybersecurity market growth through 2021, compared to the 8–10 percent projected by several industry analysts." Steve Morgan, "Global Cybersecurity Spending Predicted To Exceed $1 Trillion From 2017–2021," *Cybercrime Magazine*, June 10, 2019, https://cybersecurityventures.com/cybersecurity-market-report/.

7   James Lewis, *Economic Impact of Cybercrime: No Slowing Down*, McAfee-CSIS report, February 2018, https://csis-website-prod.s3.amazonaws.com/s3fs-public/publication/economic-impact-cybercrime.pdf.

8   "A New Kind of Insanity: The Risk of Diminishing Returns in Cybersecurity," Lumen, March 28, 2018, https://blog.lumen.com/a-new-kind-of-insanity-the-risk-of-diminishing-returns-in-cybersecurity/.

9   L.A. Gordon and M.P. Loeb, "The Economics of Information Security Investment," *ACM Transactions on Information and System Security* 5, no. 4 (November 2002): 438–457.

10 See W. Brian Arthur, "Increasing Returns and the New World of Business," *Harvard Business Review* (July–August, 1996), 101–109.

It is only prudent to anticipate that returns from investing in contemporary cybersecurity will not keep pace with changing threats as they become increasingly sophisticated and gain more opportunities to wreak mischief.

Cyber threats themselves range from lone-wolf cybercriminals to great-power rivals waging cyberwar. Although cybercrime is increasingly harmful, few if any cybercriminals have the means to compromise the encryption of communications. But great powers do.

Russia, although weak in many of the costlier sorts of power, such as conventional military forces, can launch devastating cyberattacks and views Western democracies as prime targets. Meanwhile, its relatively modest reliance on networked data makes it hard to deter by the threat of retaliation-in-kind. Recent disclosures about Russian penetration of important U.S. government networks have shattered faith in U.S. cybersecurity and shown that the ingenuity of Russian offensive operatives surpasses that of U.S. defenders. The ability of Moscow's hackers to smuggle malicious code undetected into U.S. government agencies via system updates of SolarWinds software indicates a dismal return on the more than $19 billion (FY2020) the federal government has invested annually on cybersecurity.[11] Clearly, Russian hackers are besting U.S. cybersecurity.

China, for its part, is developing advanced information technology in order to compete with the United States economically, as well as to challenge it militarily in the vital Indo-Pacific region. Quantum computing, if practical, could offer intelligence advantages to the side that can use it to break many of today's encryption keys within a reasonable time. China might thus parlay a lead in quantum technologies into superiority in cybersecurity. Both China and the United States are quite vulnerable to cyberattack by virtue of their economic dependence on data networks. Consequently, a tacit mutual deterrence is in place.[12] But will this hold if China achieves superiority in offensive *and* defensive capabilities in cyberspace as a result of its technological investments? The Chinese state and its associated technology companies are treating quantum technology as a particularly high priority among them. Even as Google and IBM race to produce useful quantum machines, China's Alibaba and Huawei are doing the same.

China is also putatively ahead of the United States in quantum communication.[13] The Chinese have demonstrated the feasibility of unbreachable quantum links through the

---

11 "Proposed Federal Spending by the U.S. Government on Cyber Security for Selected Government Agencies from FY 2020 to FY 2021", Statista, February 2020, https://www.statista.com/statistics/737504/us-fed-gov-it-cyber-security-fy-budget/.

12 See David C. Gompert and Phillip C. Saunders, *The Paradox of Power* (Washington, DC: NDU Press, 2011).

13 See David C. Gompert, "Spin-On: How the U.S. Can Meet China's Technological Challenge," *Survival* 62 (2020).

vacuum of space, over short distances in the air,[14] and at increasing distances via very clean fiber-optic lines.

In crafting cybersecurity strategy, it helps to distinguish applications that need high-end security from the mass of users that will only pay for general security. Highly sophisticated threats, such as those from the Russians and Chinese, target critical and well-protected networks, such as those supporting national security, other sensitive government functions, critical infrastructure, key sectors, and vital financial systems. Such attacks can, if successful, have grave effects. At the same time, an increasingly large volume of cybercrime by non-state hackers may undermine use of and faith in less-protected Internet-based commercial and public networks, albeit with less significant case-by-case effects.

At present, only foreign cyberpowers are both able and motivated to attack well-protected networks of importance to U.S. and allied national security. By contrast, common hackers are both constrained and inclined to target less-protected mass-use networks. It must be noted that high-end cybersecurity relies much more on encryption standing up to attack than does mass cybersecurity, where the presence of encryption suffices to send hackers looking elsewhere for weaknesses, notably by hijacking users' computers and then reading traffic from the inside.

B. The Role of Quantum Communication

Adequate cybersecurity could become more expensive yet still be found wanting—unless new options are developed.

Quantum physics offers one such option to make keeping secrets easier. Encryption, which is how secrets are kept, comes in two types: symmetric and asymmetric. Symmetric encryption uses the same key to encrypt and decrypt; it does so very efficiently, but it requires that key to be shared—and in that process, the key is vulnerable to being intercepted. This can be a problem if one party to a conversation can only be reached through an insecure channel. Asymmetric encryption uses one key to encrypt and another one to decrypt. Because the decryption key never leaves home, it is secure, provided that the decryption key (the "private" key) cannot be inferred from the encryption key (the "public" key). Once asymmetric encryption is used to pass the keys for symmetric encryption, the latter can be used to protect communications.

That said, quantum technology can cut both ways in respect to cybersecurity: whereas *quantum communication* could bolster cybersecurity, *quantum computing* could worsen it. In the words of a leading cybersecurity analyst, attacks on cryptography

---

14   Juan Yin et al., "Entanglement-Based Secure Quantum Cryptography over 1,120 Kilometres," *Nature*, June 15, 2020, https://www.nature.com/articles/s41586-020-2401-y.

systems "always get better; they never get worse."[15] This will be especially true when quantum computing becomes available. Since 1994[16] it has been known that a quantum computer could factor prime numbers in polynomial time, rather than the prohibitive exponential time currently required.[17] The difficulty of factoring numbers into primes is the current basis for believing that asymmetric encryption is secure. If someone discovers how to make factoring simpler, encryption-key security can be compromised. Against this threat, the cryptographic community is developing quantum-resistant algorithms (such as lattice-based cryptography and super-singular isogeny Diffie-Hellman key exchange), but one of the dangers of relying on these is that the security of such systems has yet to be proven. While no such quantum-computing threats are known to endanger symmetric encryption, the latter still has to solve the problem of exchanging keys securely.

This is where quantum communication comes in, specifically for quantum-key distribution (QKD). Thanks to a key feature of quantum physics, particle entanglement, it is possible to *prove* that a message was not intercepted. Martin Giles notes: "The beauty of qubits from a cybersecurity perspective is that if a hacker tries to observe qubits in transit, their super-fragile state causes them to collapse into 1 or 0 digital bits."[18] QKD, in turn, would have two parties use quantum encryption to exchange symmetric encryption keys. If the exchange was tapped, the parties would instantly know and try again. If it was untapped, the parties could use the keys with confidence.

Although prototype QKD systems have been engineered, the bandwidth along all these channels is low: though this is not a problem for exchanging keys or short, highly classified messages, it is a problem for broadband applications. Another challenge is that distances of practical quantum communication (for QKD) are limited to tens of kilometers. Repeating delicate qubits is much harder than repeating digital bits. Although scientists have shown that quantum repeaters are theoretically possible and have developed the various steps that comprise them, they have not yet produced a working prototype.[19] China has performed long-range line-of-site transmission through

15   Bruce Schneier, "New Attack on AES," *Schneier on Security*, August 18, 2011, https://www.schneier.com/blog/archives/2011/08/new_attack_on_a_1.html.

16   P.W. Shor, "Algorithms for Quantum Computation: Discrete Logarithms and Factoring," *Proceedings 35th Annual Symposium on Foundations of Computer Science*, 124–134 (IEEE Computer Society Press, 1994).

17   Researchers have made impressive progress at developing quantum computing since 1994. That said, they have yet to develop a practical instantiation of a computer that can efficiently crack prime numbers (exchange with author, May 2020). And researchers at the Princeton Quantum Initiative believe that codebreaking with quantum computing will not be feasible anytime soon (exchange with author, May 2020).

18   Martin Giles, "Explainer: What is Quantum Communication," *MIT Technology Review*, February 14, 2019, https://www.technologyreview.com/2019/02/14/103409/what-is-quantum-communications/.

19   According to a member of the Princeton Quantum Initiative: "All of the steps involved [in qubit repeating] have been demonstrated experimentally at a proof-of-concept level: people have demonstrated spin-photon entanglement, two-photon interference, remote entanglement distribution, quantum teleportation, and entanglement distillation. They just have not demonstrated a platform that is capable of break-even repeater networks to get to long distances. This is sort of analogous to current quantum computers—people have demonstrated quantum error correction, but only barely break-even, and not in a way that scales to large systems" (exchange with author, November 2020).

space via its Micius quantum-communication satellite,[20] but the practical applications are unclear. The United States has yet to deploy a quantum-communication satellite. The Chinese are also working on drones as quantum-communication nodes, but their ranges are too short to be of much utility.[21]

Granted, quantum communication alone will not guarantee cybersecurity. As a "system-of-systems" problem, cybersecurity requires a vast variety of tasks be done right: e.g., determining authorized users; authenticating their identity; protecting the integrity of applications and data; preventing unauthorized altering of hardware and software; and protecting the inviolability of channels. When hackers defeat these measures, organizations must detect their presence, ascertain and contain their effects, and patch holes that let them enter. They may have to develop plans to work around and recover from attacks. Even if communication links are protected by quantum communication, digital platforms could still be insecure. Other vulnerabilities include poor access control; ill-advised protocols; malware-laden computers, clients, servers or routers; and vulnerable supply chains.

Nevertheless, quantum communication can be a game-changer, at a minimum for safeguarding encryption. Thus the key question is how to proceed strategically.

2. STRATEGIC OPTIONS FOR A QUANTUM INTERNET

The expansion of remote work and associated network use is also occurring in more critical endeavors, such as government and proprietary corporate business. Such remote work challenges cybersecurity at the mass end of the spectrum. Yet the quantum communication technologies under current development are geared to problems at the high end where the quality of encryption is crucial: these require specialized hardware, which remains expensive because of the intricate engineering required to keep error rates for qubits low enough to allow reliably readable results.

This creates a dilemma: the problem of cybersecurity is growing for mass applications, but quantum technology, at least for now, offers relief mainly at the high end. This presents a conundrum: whether, how, and how fast to steer quantum communication development to address both high and mass segments of the Internet. At its core, this is about how markets and governments affect the progress of technology. Markets pull technology, and governments push it.

---

20   See, for instance, Karen Kwon, "China Reaches New Milestone in Space-Based Quantum Communications: The Nation's Micius Satellite Successfully Established an Ultrasecure Link between Two Ground Stations Separated by More Than 1,000 Kilometers," June 25, 2020, *Scientific American*, https://www.scientificamerican.com/article/china-reaches-new-milestone-in-space-based-quantum-communications/.

21   Anil Ananthaswamy, "The Quantum Internet Is Emerging, One Experiment at a Time," *Scientific American*, June 19, 2019, https://www.scientificamerican.com/article/the-quantum-internet-is-emerging-one-experiment-at-a-time/.

Consider three ways to approach this conundrum:

- Accept that quantum communication technology cannot serve the mass market.
- Push the technology in general but let it find its own markets.
- Encourage the technology to address the mass market.

Below, we take each in turn.

A. Accept That the Technology Cannot Serve the Mass Market

Not every technology benefits a mass user base, and not every technology that benefits high-end users has only high-end potential. Sixty-four years after Sputnik, for instance, rocketry is still the province of countries, corporations, and a few rich individuals (e.g., Elon Musk and SpaceX, Jeff Bezos and Blue Origin, Richard Branson and Virgin Galactic). Yet the orbits that rockets have opened up for use have brought accurate weather forecasting, the Global Positioning System (GPS), and satellite television to the masses.

By contrast, digital technology went down-market towards mass use almost from its inception. In the 1980s, the emergence of fiber optics and personal computers resulted in a mass shift toward distributed processing and broadband data networking. The most important markets for this integration of computing and telecommunications were large decentralized corporations and growing numbers of individuals. Although government funding provided some impetus, it was the ballooning revenues from civilian demand that provided the fuel for research and development that gave life to the digital revolution. Government clients, even the military and intelligence community, lagged at first but eventually climbed aboard, resulting in specialized sensitive networks. Cybersecurity, unfortunately, was an afterthought.

Quantum communication could follow a very different path: because its principal benefit is to bolster encryption, the most obvious application is to secure sensitive domains from sophisticated foreign-power threats. True, countering high-end threats can benefit everyone: we all rely on national security, financial, and critical infrastructure systems. But QKD is not needed for the security of mass networks.

Conversely, even if QKD moves "down-market," it is unclear whether an advance in encryption technologies can improve cybersecurity all that much (even as the reverse is true: advances in decryption generally harm cybersecurity). Two wise cybersecurity experts, Ross Anderson and Bruce Schneier, began their careers in cryptography with a belief that better cryptography was needed to improve cybersecurity. Both concluded that while good cryptography mattered,[22] better cybersecurity was more

---

[22] Even after reaching that conclusion, Bruce Schneier co-designed Blowfish, a symmetric encryption algorithm that was the runner-up in National Institute for Standards and Technology's competition to develop a new symmetric key encryption standard. See Bruce Schneier, "The Blowfish Encryption Algorithm," *Schneier on Security*, accessed April 8, 2021, https://www.schneier.com/academic/blowfish/.

likely to emerge from a much broader understanding of security *per se* and a thorough adjustment in the incentives that decision-makers face when weighing cybersecurity decisions.[23]

Indeed, what kind of relief can technological development in general provide to cybersecurity? Start with the premise that all cybersecurity faults originate in human behavior. True as that may be, the primary implications—whether that cyber insecurity is deeply rooted in human nature and is hence ineradicable, or that cybersecurity is primarily sought through improving human behavior—do not necessarily follow. The most cost-effective path forward in such cases may involve not improving humans but establishing systems that prevent or mitigate the consequences of bad human decisions (or user interfaces that check potentially harmful but reflexive acts). Almost all automobile accidents, for instance, stem from human error. Yet between 1966 and 2014, in the United States, the number of fatalities per vehicle mile traveled fell by a factor of five (from 55 to 11 per billion miles traveled).[24] Are U.S. drivers five times better today (apart from declines in drunken and adolescent driving)? Or is the reduction more a result of better cars (seat belts, air bags, warning systems, frame integrity), better roads (freeways), and more efficient emergency medical services?

Similarly, even if better *user* choices help, the choices made by systems administrators and their bosses may help even more. And better technologies should not be confused with better techniques. Both technology and technique involve know-how. We think of technology as explicit, with universal properties that are globally applicable rather than the solution of a problem that varies by circumstance; it is thus capable of being transferred. Techniques belong to those who have mastered them and are thus far harder to transfer. There is very little "once-and-for-all" in the field of cybersecurity. Measures beget countermeasures, which beget counter-countermeasures, and so on. By contrast, quantum entails the mastery of new physical principles.

Artificial intelligence (AI) has been touted as a technology that can both improve and harm cybersecurity. Results from the Defense Advanced Research Projects Agency (DARPA) Grand Challenge program indicate that AI can spot software vulnerabilities better than humans can.[25] But that cuts both ways. AI can help vendors build more secure software. But AI can also help state-sponsored actors find some software or network vulnerability first. It is unclear whether accelerating the rate by which both

---

23   See, for instance, Ross Anderson, "Why Cryptosystems Fail," paper presented at the Association for Computing Machinery Conference on Computer and Communications Security (Fairfax, VA, November 1993), https://www.cl.cam.ac.uk/~rja14/Papers/wcf.pdf.

24   Wikipedia, "Motor Vehicle Fatality Rate in U.S. by Year," last modified April 5, 2021, https://en.wikipedia.org/wiki/Motor_vehicle_fatality_rate_in_U.S._by_year. See also the detailed statistics from United States Department of Transportation, "Recent NCSA Publications," accessed April 8, 2021, https://crashstats.nhtsa.dot.gov/.

25   See, for instance, David Brumley, "Mayhem, the Machine That Finds Software Vulnerabilities, Then Patches Them," *IEEE Spectrum*, January 29, 2019, https://spectrum.ieee.org/computing/software/mayhem-the-machine-that-finds-software-vulnerabilities-then-patches-them.

sides discover vulnerabilities will improve cybersecurity.[26] To the extent, however, that AI means machine learning, *and* that machine learning is used to spot network anomalies indicative of an intrusion, there are grounds for believing that AI will improve cybersecurity; if nothing else, it should improve configuration and patch management. But if hacking works by playing against expectation—and especially if hackers have access to AI that they can practice against to improve their ability to work undetected beneath some noise level—there may simply not be reliable corpora of abnormal network behavior to work with.

Furthermore, if a consequence of pursuing a technology is to hasten its adoption by others—as has been the case with digital technologies—second thoughts about the wisdom of doing so may be in order. The most important "other" is China, which, as noted, has actively pursued quantum communication, motivated by a belief in the power of U.S. intelligence agencies to ferret out secrets. But China no longer depends on U.S. technology to bootstrap such efforts, and so a U.S. failure to pursue such a technology would offer no help vis-à-vis China.

Many threat actors, however, cannot finance quantum communication advances or even exploit them at current prices. If further advances in quantum communication become useful, though, then U.S. efforts to thwart hackers by hacking them (e.g., "persistent engagement") might be that much harder. This is an example of what has been labeled the "cybersecurity dilemma."[27] However, an opposing argument can also be made. Hackers are a group that, once burned, might become wise to such efforts and therefore able to resist[28] without the help of quantum communication to mask their doings. Those hackers with less sophistication or resources may not be able or willing to take advantage of even tomorrow's quantum communication. Thus its advent would have little effect on their vulnerability to the various tools of "persistent engagement." Similar conclusions may apply more broadly. Although groups such as drug cartels also have an interest in encrypted communications, commercial technologies carefully implemented (e.g., Signal, Telegram) may suffice, because they are trying to evade national police agencies, not national intelligence agencies. Quantum communication, at this point, is more suited to network architectures with

---

26 This touches on a long-running debate over whether vulnerabilities are common (in which case, such an acceleration would not make much difference) or sparse (in which case, it would). Ross Anderson ("Security in Open versus Closed Systems: The Dance of Boltzmann, Coase, and Moore," Open Source Software: Economics, Law and Policy, IDEI Presentation, Toulouse, France, June 20–21, 2002. https://www.helpnetsecurity.com/2002/07/09/security-in-open-versus-closed-systems-the-dance-of-boltzmann-coase-and-moore/) thinks that neither attackers nor defenders gain a definitive advantage from open source software. However, empirical work by Andrew Ozment and Stuart E. Schechter ("Milk or Wine: Does Software Security Improve with Age?" Report, Usenix, 2006. http://www.usenix.org/legacy/event/sec06/tech/full_papers/ozment/ozment.pdf) suggests that depletion is possible, and hence, AI would correlate with greater cybersecurity.

27 See Ben Buchanan, *The Cybersecurity Dilemma: Hacking, Trust and Fear Between Nations* (Oxford: Oxford University Press, 2017).

28 This is not to say that "persistent engagement" is worthless. Forcing threat hackers to build a more robust attack infrastructure or cover their tracks more carefully detracts from their overall efforts.

few (albeit long-haul) nodes rather those with many nodes, because nodes themselves create opportunities for interception.

B. Let the Technology Find Its Own Markets

The history of technology is rife with instances in which a new capability initially seems irrelevant to the problems of everyday individuals but then—as it becomes more reliable, easier to use, and, especially, cheaper—becomes widespread and benefits almost everyone directly. Automobiles underwent such a shift in the United States from the late 1890s to the early 1920s. Computers did so from the 1950s-era mainframe that only a few organizations could afford to own (and, as importantly, service) to the early 1980s-era personal computers. Conversely, the benefits of many important technologies, notably aviation, filtered down to the masses only through their uptake by organizations (e.g., airlines). And the link between technologies that support national security and those that benefit the population at large is highly indirect.

Will quantum technologies filter to the masses directly, or will their benefits be realized only by and through those who can afford it, such as governments and banks? It is hard to be optimistic that quantum computation and communication will take the direction their predecessors did. A disproportionate share of the technological advances over the last 50 years has come from the ability to manipulate matter at an increasingly small scale. The march of semiconductor performance (known as Moore's Law) has resulted in large part from the constant shrinkage of integrated circuit size from 10 microns (circa 1970) to .007 microns (circa 2021). Sequencing a human genome, which cost roughly $100 million in 2000, now costs under $1000.[29] Similar advances have affected nanomaterial structures. By contrast, technological progress in the preceding 50 years (1920–1970) resulted from the ability to scale up processes so that products once manufactured in factories sized to fill regional needs were now supplied by factories scaled to global markets.

Quantum technologies arise from advances in working at ever-more-precise process control; they are very sensitive to environmental conditions. Progress requires erasing or compensating for all sources of extraneous noise (i.e., unwanted signal). It is a technology which, in spirit, is similar to those which enable precision ballistics.[30] These are not the sorts of technologies that allow rapid advances in scale—at least not in comparison to when a single process (e.g., photolithography) achieves great economics by producing an ever-larger number of products per unit (e.g., transistors per square inch of wafer).

---

[29] National Human Genome Research Institute, "The Cost of Sequencing a Human Genome," National Human Genome Research Institute website, https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

[30] See Donald MacKenzie, *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance* (Cambridge, MA: MIT Press, 1990).

Ironically, insofar as miniaturization has now put cheap microprocessors not only in computers but in computer peripherals and internetted devices (e.g., light bulbs), it has complicated security. There are now many more places for malware to hide.

C. Push Quantum Communication Down-Market

The last option—deliberately encouraging the progression of quantum technology down-market—would reflect the judgment that government support and inducements work and that otherwise cybersecurity for mass use would not be helped. But two questions immediately arise: where would the technology be pushed, and how?

Although quantum communication can improve cybersecurity against the threat of interception, the interception at issue comes from tapping links rather than nodes (such as client machines and routers). Links come in two main types: wired and wireless. Wireline tapping requires operating some device at or very close to the line. The physical proximity and surreptitiousness required for wireline tapping makes it the province of governments and hence of limited usefulness to hackers of everyday users. Although tapping trunk lines can be and is done, the use of quantum channels for trunk lines would greatly exceed the bandwidth currently available to quantum communication (even if it suffices for QKD).

But wireless tapping is easier, since the distance between the tap and the channel need only be comparable to that between two nodes. It does not require a state apparatus to pull it off; infecting devices that the user already has could suffice (even if exfiltrating data undetected takes additional work). With more and more devices capable of being networked via Bluetooth, Wi-Fi, or perhaps 5G, avoiding such interception may become an increasing component of cybersecurity for the home, office, or factory.

Therein lies a dilemma. Internet-of-things (IoT) devices tend to be insecure because they often transmit in the clear. They abjure encryption because generating the processor cycles needed for encryption and decryption can be burdensome for cheap low-power processors. Meanwhile, quantum devices are sizeable and must be highly sensitive to the ambient environment to work reliably. Of course, so were early computers, as those who remember carefully filtered air-conditioned computer rooms will understand. Computers did not evolve to serve personal needs until integrated circuits were developed. If—and this is a huge if—there were ways to reduce quantum communication's read/write capabilities to integrated circuit form, it may be possible to embed quantum communication into any and all radio-frequency (RF) processing chips. As a bonus, because such devices could detect the presence of interception, they could also be used as high-fidelity sensors for listening devices. But none of this will happen soon.

In the meantime, there are other ways to push quantum technology toward mass cybersecurity. They include introducing it into cloud computing, particularly in server-to-server communication, and perhaps developing quantum-as-a-service. But quantum communication must first be proven cost-effective on its own terms before having additional demands thrust on its technological development.

3. RECOMMENDATIONS

Of these strategic alternatives, the authors lean toward the third, which would call for a public-private approach to make the technology robust and push it down-market. This would direct its benefits to those whom we expect to stay online even after the Covid-19 pandemic winds down. At the same time, the economic means to exploit quantum communication for the sake of mass cybersecurity must come mainly from markets themselves: in research and development, to advance the technology, notably to overcome the distance and bandwidth problems; in capital, to augment the existing Internet with quantum links; and in revenue-generating demand, for better security from eager users of every sort. If quantum communication is sufficiently promising, market-demand signals should augment government initiatives to introduce and spread this technology's use and value.

We recommend this strategy for several reasons. Even if highly sensitive links are made more secure, the increased cyber vulnerability of mass networks is, broadly conceived, a national security problem that cannot be ignored lest economic losses mount while information leaks voluminously. Citizens will lose confidence in their access to trustworthy information, in their government's ability to safeguard it, and in the reliability of elections and health of democracy itself.

To implement this public-private strategy, we recommend several specific steps:

- The U.S. government (notably the Department of Energy, the Department of Homeland Security, and the Department of Defense), allied governments, leading information-technology companies, and major universities should jointly commit to developing and deploying quantum communication.
- Concerted, yet still competitive, efforts should be made to overcome range and bandwidth obstacles. A combination of government-funded and corporate research and development investment is needed. Similarly, concerted engineering efforts should be made on cost reduction, especially if the technology can be driven towards a chip-level orientation.
- High priority should be given to domains of direct importance to national security.

- High priority should also be given to the protection of intellectual property rights, coupled with widespread licensing.
- U.S.-allied partnerships should be promoted; European quantum work (e.g., at Delft University) is world-class, as is reflected in current partnerships. Indeed, one of the better venues for such collaboration would be NATO, which already includes cybersecurity among its missions.

Although governments cannot insist that private technology companies team with others who may compete with them, it can galvanize teaming. With its proven capacity for facilitating cooperation in sensitive defense and intelligence affairs, NATO (with arrangements to include Japan and certain other partners) is a natural place to start.

REFERENCES

Accenture. *Ninth Annual Cost of Cybercrime Study*. 2019. https://www.accenture.com/_acnmedia/pdf-96/accenture-2019-cost-of-cybercrime-study-final.pdf.

Ananthaswamy, Anil. "The Quantum Internet Is Emerging, One Experiment at a Time." *Scientific American*. June 19, 2019. https://www.scientificamerican.com/article/the-quantum-internet-is-emerging-one-experiment-at-a-time/.

Anderson, Ross. "Security in Open versus Closed Systems: The Dance of Boltzmann, Coase, and Moore." Open Source Software: Economics, Law and Policy, IDEI Presentation, Toulouse, France, June 20–21, 2002. https://www.helpnetsecurity.com/2002/07/09/security-in-open-versus-closed-systems-the-dance-of-boltzmann-coase-and-moore/.

Anderson, Ross. "Why Cryptosystems Fail." Paper presented at the Association for Computing Machinery Conference on Computer and Communications Security, Fairfax, VA, November 1993. https://www.cl.cam.ac.uk/~rja14/Papers/wcf.pdf.

Arthur, W. Brian. "Increasing Returns and the New World of Business." *Harvard Business Review* (July–August 1996): 101–109.

Beech, Mark. "COVID-19 Pushes up Internet Use 70% and Streaming More Than 12%, First Figures Reveal." *Forbes*. March 25, 2020. https://www.forbes.com/sites/markbeech/2020/03/25/covid-19-pushes-up-internet-use-70-streaming-more-than-12-first-figures-reveal/.

Brumley, David. "Mayhem, the Machine That Finds Software Vulnerabilities, Then Patches Them." *IEEE Spectrum*. January 29, 2019. https://spectrum.ieee.org/computing/software/mayhem-the-machine-that-finds-software-vulnerabilities-then-patches-them.

Buchanan, Ben. *The Cybersecurity Dilemma: Hacking, Trust and Fear Between Nations*. Oxford: Oxford University Press, 2017.

Cimpanu, Catalin. "FBI Says Cybercrime Reports Quadrupled during COVID-19 Pandemic." *ZD Net*. April 18, 2020. https://www.zdnet.com/article/fbi-says-cybercrime-reports-quadrupled-during-covid-19-pandemic/.

Cohen, Jason. "Data Usage Has Increased 47 Percent During COVID-19 Quarantine." *PCMag*. June 5, 2020. https://www.pcmag.com/news/data-usage-has-increased-47-percent-during-covid-19-quarantine.

Gartner. "Gartner Forecasts Worldwide Security and Risk Management Spending Growth to Slow but Remain Positive in 2020." Gartner Newsroom, June 17, 2020. https://www.gartner.com/en/newsroom/press-releases/2020-06-17-gartner-forecasts-worldwide-security-and-risk-managem.

Gerwitz, David. "COVID Cybercrime: 10 Disturbing Statistics to Keep You Awake Tonight." *ZDNet*. September 14, 2020. https://www.zdnet.com/article/ten-disturbing-coronavirus-related-cybercrime-statistics-to-keep-you-awake-tonight/.

Giles, Martin. "Explainer: What is Quantum Communication." *MIT Technology Review*. February 14, 2019. https://www.technologyreview.com/2019/02/14/103409/what-is-quantum-communications/.

Gompert, David C. "Spin-On: How the US Can Meet China's Technological Challenge." *Survival* 62, no. 3 (2020): 115–130. DOI: 10.1080/00396338.2020.1763617.

Gompert, David C., and Phillip C. Saunders. *The Paradox of Power*. Washington, DC: NDU Press, 2011.

Gordon, L.A., and M.P. Loeb. "The Economics of Information Security Investment." *ACM Transactions on Information and System Security* 5, no. 4 (November 2002): 438–457.

Grossman Group. "Nearly Half of Employees Now Working from Home Want to Stay Remote, Study Finds." PR Newswire. May 14, 2020. https://www.prnewswire.com/news-releases/nearly-half-of-employees-now-working-from-home-want-to-stay-remote-study-finds-301059220.html.

Kwon, Karen, "China Reaches New Milestone in Space-Based Quantum Communications: The Nation's Micius Satellite Successfully Established an Ultrasecure Link between Two Ground Stations Separated by More Than 1,000 Kilometers," *Scientific American*, June 25, 2020; https://www.scientificamerican.com/article/china-reaches-new-milestone-in-space-based-quantum-communications/.

Lewis, James. *Economic Impact of Cybercrime: No Slowing Down*. McAfee-CSIS report. February 2018. https://csis-website-prod.s3.amazonaws.com/s3fs-public/publication/economic-impact-cybercrime.pdf.

Lumen. "A New Kind of Insanity: The Risk of Diminishing Returns in Cybersecurity." Lumen website. March 28, 2018. https://blog.lumen.com/a-new-kind-of-insanity-the-risk-of-diminishing-returns-in-cybersecurity/.

MacKenzie, Donald. *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance*. Cambridge, MA: MIT Press, 1990.

Morgan, Steve. "Global Cybersecurity Spending Predicted To Exceed $1 Trillion From 2017–2021." *Cybercrime Magazine*. June 10, 2019. https://cybersecurityventures.com/cybersecurity-market-report/.

National Human Genome Research Institute. "The Cost of Sequencing a Human Genome." National Human Genome Research Institute website. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

Newman, Lily Hay. "The NSA Warns That Russia Is Attacking Remote Work Platforms." *Wired*. December 7, 2020. https://www.wired.com/story/nsa-warns-russia-attacking-vmware-remote-work-platforms/.

Ozment, Andrew, and Stuart E. Schechter. "Milk or Wine: Does Software Security Improve with Age?" Report, Usenix. 2006. http://www.usenix.org/legacy/event/sec06/tech/full_papers/ozment/ozment.pdf.

Reuters. "Edited Transcript of BLK.N Earnings Conference Call or Presentation 13-Oct-20 12:30pm GMT." Yahoo Lifestyle. October 13, 2020. https://www.yahoo.com/lifestyle/edited-transcript-blk-n-earnings-123000634.html.

Schneider, Troy. "Nearly 50% of Pentagon Workers Still Teleworking." *FCW (Federal Computer Week)*. September 17, 2020. https://fcw.com/articles/2020/09/17/pentagon-telework-fifty-percent.aspx.

Schneier, Bruce. "New Attack on AES." *Schneier on Security*. August 18, 2011. https://www.schneier.com/blog/archives/2011/08/new_attack_on_a_1.html.

Schneier, Bruce. "The Blowfish Encryption Algorithm." *Schneier on Security*. Accessed April 8, 2021. https://www.schneier.com/academic/blowfish/.

Shor, P.W. "Algorithms for Quantum Computation: Discrete Logarithms and Factoring." In *Proceedings 35th Annual Symposium on Foundations of Computer Science*, 124–134. IEEE Computer Society Press, 1994.

Statista. "Proposed Federal Spending by the U.S. Government on Cyber Security for Selected Government Agencies from FY 2020 to FY 2021." February 2020. https://www.statista.com/statistics/737504/us-fed-gov-it-cyber-security-fy-budget/.

United States Department of Transportation. "Recent NCSA Publications." Accessed April 8, 2021. https://crashstats.nhtsa.dot.gov/.

U.S. Patent and Trademark Office, Patent Public Advisory Committee. *Annual Report 2020*. October 30, 2020. https://www.uspto.gov/sites/default/files/documents/PPAC_2020_Annual_Report.pdf.

Vogels, Emily, Andrew Perrin, Lee Rainie, and Monica Anderson. "53% of Americans Say the Internet Has Been Essential during the COVID-19 Outbreak: Americans with Lower Incomes Are Particularly Likely to Have Concerns Related to the Digital Divide and the Digital 'Homework Gap.'" Pew Research Center. April 30, 2020. https://www.pewresearch.org/internet/2020/04/30/53-of-americans-say-the-internet-has-been-essential-during-the-covid-19-outbreak/.

Yin, Juan, et al. "Entanglement-Based Secure Quantum Cryptography over 1,120 Kilometres." *Nature*. June 15, 2020. https://www.nature.com/articles/s41586-020-2401-y.

Wikipedia. "Motor Vehicle Fatality Rate in U.S. by Year." Last modified April 5, 2021. https://en.wikipedia.org/wiki/Motor_vehicle_fatality_rate_in_U.S._by_year.

BIOGRAPHIES

Editors

Taťána Jančárková is a researcher in the Law Branch at NATO CCDCOE, where she works on the application of international law to cyberspace operations (the Cyber Law Toolkit project) and cyber security-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 negotiating EU cyber legislation and cyber defence-related cooperation with NATO and the 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.

Maj. **Gábor Visky** is a researcher in the Technology Branch at NATO CCDCOE, where his main field of expertise is industrial control systems. His previous assignments include 15 years of designing hardware and software for embedded control systems and researching their vulnerabilities through reverse engineering. Major Visky holds an MSc in information engineering with a major in industrial measurement and a BSc in the field of telecommunications.

Philippe Mitsuya Zotz has a background in media studies and social informatics, and holds a position at the Department for Communications and Information Systems at the Luxembourg Armed Forces Staff. He was seconded to the NATO CCDCOE and joined the Centre's Strategy branch in 2020. Apart from cyber defence, Mr Zotz' fields of interest also include CIS security, online privacy and digitalisation.

Authors

Pietro Baroni is full professor of computer science and engineering at the Department of Information Engineering of the University of Brescia, Italy. He is author of more than 130 scientific papers in the area of artificial intelligence and knowledge-based systems, with a main focus on theory and applications of computational argumentation. He was a founding member of the Steering Committee of the COMMA (*Computational*

Models of Argument) conference series and is currently one of the editors-in-chief of the *Argument & Computation* journal.

Dr **Jason Blessing** is a DAAD Post-Doctoral Fellow with the United States, Europe, and World Order Program in the Foreign Policy Institute at the Johns Hopkins University School of Advanced International Studies (SAIS) in Washington DC. He is also a Consulting Fellow for the Cyber, Space and Future Conflict Programme at the International Institute for Strategic Studies (IISS), where he writes on cyber military maturity and is developing a new methodology for assessing military cyber forces. Jason is a former USIP-Minerva Peace and Security Scholar, a position jointly funded by the United States Institute of Peace and the US Department of Defense Minerva Research Initiative. He holds a PhD in political science from the Maxwell School of Citizenship and Public Affairs at Syracuse University, an MA in political science from Virginia Tech, and a BA in government from The College of William & Mary. Prior to his doctoral degree, Jason was a licensed Fraud Operations Analyst for a US-based brokerage firm, where he identified, mitigated, and resolved cases of financial fraud based on technical compromise data and client behavioural patterns.

Prof. **Thomas Brandstetter** is an OT cyber security expert, with more than 20 years of diverse experience in multiple technical and management roles. He is co-founder and managing director of Limes Security, a major European OT cyber security company, Professor for IT Security at University of Applied Sciences, St. Poelten, and Honorary Professor for Cyber Security at DeMontfort University. He also serves as review board member for Black Hat and teaches control system security classes for SANS. His past noteworthy achievements include being the incident handler for the Stuxnet malware at Siemens, as well as being the founder of the Siemens ProductCERT.

Dennis Broeders is Full Professor of Global Security and Technology and Senior Fellow of The Hague Program for Cyber Norms at the Institute of Security and Global Affairs (ISGA) at Leiden University, the Netherlands. His research and teaching broadly explores the interaction between security, technology and policy, with a specific interest in international cyber security governance. He is the author of the book *The Public Core of the Internet* (2015). He currently also serves as a member of the Dutch delegation to the UN Group of Governmental Experts on international information security (2019–2021) as an academic adviser.

Tim Casey is a senior Information Security Risk Manager with Intel Corp., where he develops and executes programmes to mitigate critical cyber risks to the corporation, such as initiating and directing Intel's wargaming, insider risk, and policy strategy programmes. He developed the threat agent ontology and library there and continues to lead a cross-organisational team of senior analysts in maturing and applying its

concepts to anticipate and manage cyber and insider threats. Mr Casey has also led some of Intel's high-impact public-private engagements, including architecting a new model for US critical infrastructure risk assessments, and helping create and pilot the NIST Cybersecurity Framework.

Federico Cerutti is the Chair of the Brescia University branch of the Italian Cybersecurity National Laboratory. In 2019, he was one of the recipients of the highly competitive Rita Levi-Montalcini personal research fellowship programme funded by the Italian Ministry of Research. Before that, he was the Academic Director of the Cardiff University Data Science Academy and Senior Lecturer at the same institution. He is an active researcher in learning and reasoning with uncertainty and cyber threat intelligence analysis. He has published over 70 peer-reviewed papers, several in top publications for artificial intelligence research.

Fabio Cristiano is a postdoctoral researcher at The Hague Program for Cyber Norms in the Institute of Security and Global Affairs (ISGA) at Leiden University, where he also teaches BA Security Studies and Executive MSc Cyber Security. His research and teaching broadly lies at the intersection of critical security studies and international relations theory, with a specific interest in automation, autonomy, and international norms in the context of cyber and information warfare.

François Delerue (PhD) is a research fellow in cyber defence and international law at the Institute for Strategic Research (Institut de recherche stratégique de l'École militaire – IRSEM, Paris, France) and a lecturer at Sciences Po. Dr Delerue is also a non-resident associate fellow at The Hague Program for Cyber Norms at Leiden University and a rapporteur on international law for the Academic Advisory Board of EU Cyber Direct. His book *Cyber Operations and International Law* was published by Cambridge University Press in February 2020.

Daniela Fogli is Professor at the Department of Information Engineering, University of Brescia, Italy. Her research interests include methods for designing complex interactive systems, meta-design, end-user development, web usability and accessibility and decision support systems. She has performed her research activity in collaboration with several scholars of different universities and research centres. She serves as senior associate editor for the *Decision Support Systems* journal, and she is chair of the steering committee of the International Symposium of End-User Development (IS-EUD). At the University of Brescia, she teaches Human-Computer Interaction and Foundations of Computer Science.

Luca Gambazzi (PhD) graduated in microengineering from the EPFL and then obtained a PhD in 2010. He has worked in the field of data analysis, processing large amounts of data in different projects in Switzerland and abroad. Since 2016, he has been employed at the Federal Office for Defence Procurement of Switzerland (armasuisse). His main tasks include infosec, solutions' evaluation in procurement processes, architecture and data protection.

Massimiliano Giacomin is full professor of Computer Science and Engineering at University of Brescia. His main research interests are in knowledge representation and automated reasoning (in particular, argumentation theory, fuzzy constraints, temporal reasoning), multi-agent systems, and knowledge-based systems. He has published more than 110 papers and organised several workshops and conferences. He was keynote speaker at PRIMA 2017, and has been invited speaker at international workshops, author and co-author of tutorials for various conferences, summer schools and workshops. In 2006, he has awarded the 'Marco Somalvico' Artificial Intelligence prize by the Italian Association for Artificial Intelligence (AI*IA). He is currently a member of the editorial board of the journal *Argument & Computation*.

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. Mr Giles has been involved with the exploitation of the internet for three decades and combines a technical background with the in-depth study of approaches by authoritarian regimes to information security to develop the analysis and prediction of the development of information warfare, including the subdomain of cyber conflict. He is the author of a number of studies on Russian theory, doctrine, and structures for engaging in information and cyber confrontation.

The Honorable **David C. Gompert** is a Distinguished Visiting Professor at the US Naval Academy and Special Advisor to Ultratech Capital Partners. He has served in numerous high US government positions, including acting Director of National Intelligence, Special Assistant to President George H. W. Bush, and Special Assistant to Secretary Henry A. Kissinger. He was a senior executive in the information technology industry. Mr Gompert has written extensively on global security, defence policy, foreign policy, and the use of information technology in national security. He currently serves on several not-for-profit boards.

Roman Graf (PhD, OSCP, CEH) is Security Delivery Associate Manager at Accenture Security; his work focuses on cyber security topics (information security, DevSecOps, pentesting). Dr Graf has contributed to the development of several European research projects, including Titanium, MAL2, Ecossian, Assets and SCAPE. He has published

widely in the area of cyber security and AI. He has supported the development of the cyber threat intelligence solution CAESAIR, serving as one of the key developers, and contributed a module to the Open Source Threat Intelligence Platform (MISP).

Francesco Gringoli is Associate Professor in Computer Networks at the University of Brescia, Italy. He received a master's degree in telecommunications engineering from the University of Padova, Italy, in 1998 and a PhD in information engineering from the University of Brescia, Italy, in 2002. His research interests include security assessment, performance evaluation and medium access control in Wireless LANs. He enjoys reverse-engineering things for research, and fun.

Jonas Grätz-Hoffmann is scientific collaborator at the Federal Department of Foreign Affairs of Switzerland (FDFA), contributing to foreign policy in the areas of cyber diplomacy and digital governance. Prior to this position, he was economic adviser at the OSCE Secretariat in Vienna. From 2015 to 2018, he held different positions at the FDFA, inter alia as the desk for the OSCE's activities in Ukraine. Mr Grätz-Hoffmann worked in different European think tanks (2008–2015), such as the German Institute for International and Security Affairs (SWP) in Berlin and the Center for Security Studies at the Swiss Federal Institute of Technology, Zurich. He holds a doctorate in political science from Frankfurt University.

Giovanni Guida (PhD) received a doctoral degree in electronic engineering (major in computer science) from the Politecnico di Milano, Italy, in 1975. Since 1991, he has been a full professor of computer science at the University of Brescia, Department of Information Engineering, where he teaches digital innovation and knowledge management. Previously, he taught at the Politecnico di Milano and the University of Udine. His present research interests include decision support systems, human-computer interaction, knowledge management, argumentation, classification, and situation understanding. He is active as an independent consultant in the fields of digital strategies, intangible asset management, and intelligent systems.

Gergő Gyebnár is a Certified Information Security Professional with over 10 years developing and managing an international cybersecurity company. Mr Gyebnár has delivered high-value projects regarding Information Sharing and Analysis Centres (ISAC), Security Operations Centres, Cyber Security Incidents Response and related R&D activities. Mr Gyebnár aims to be a leader with deep knowledge in cloud and ICS security able to navigate his company in complex and dynamic environments. He studied at the National University of Public Service and soon after founded Black Cell Ltd.

Charles Harry (PhD) is a senior leader, practitioner and researcher with over 20 years in intelligence and cyber operations. Dr Harry is Director of Operations at the Maryland Global Initiative in Cybersecurity (MaGIC), an associate research professor in the School of Public Policy, University of Maryland, and a senior research associate at the Center for International and Security Studies at Maryland (CISSM). Dr Harry facilitates and promotes external engagement and interdisciplinary research across the university and is often called to speak to international and national audiences on a range of cybersecurity issues. He is part of Bain and Company's External Advisors Network and is an active consultant to a wide range of public and private organisations. Prior to his appointment at the university, he spent over two decades in the defence and intelligence communities.

Kim Hartmann serves as Cyber and Information Technology Director at the Conflict Studies Research Centre, UK. She is a senior consultant and researcher 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, IT forensics, data and privacy protection, and secure software development. She has advised private, military and government organisations for over a decade on the assessment of cyber incidents, the integration of security in existing technologies, as well as the protection of networks and software.

Kimmo Heinäaro works as a security researcher at the NATO CCDCOE. His main research task focuses on cyber-physical systems and the ICS hardware related to cyber exercises. Other topics of interest include security of industrial protocols and SDR-based approaches to pen-testing RF protocols.

Ryan Hohimer is the inventor, co-founder and CTO of DarkLight, Inc. Mr Hohimer has combined modelling techniques from Object Oriented Programming and Knowledge Representation & Reasoning concepts to create a domain agnostic artificial intelligence system based on semantic technologies. He has a diverse background that has provided him with an experiential toolkit to address very challenging technical problems ranging from the many issues of cyber security to understanding the motivations and intents of terrorists. As co-chair of the OASIS Threat Actor Context Technical Committee, focus is being placed on the semantic interoperability and exchange of a cyber adversary's context.

Dr **Benjamin M. Jensen** holds a dual appointment as a professor at the School of Advanced Warfighting, Marine Corps University, where he runs the Future War Research Program, and as a scholar in residence at American University, School of International Service. He is also a senior fellow at the Atlantic Council. Dr Jensen

has written five books on military organisations, disruptive technology and strategy, including the co-authored *Cyber Strategy: The Evolving Character of Power and Coercion* (Oxford 2018) and *Military Strategy in the 21ˢᵗ Century: People, Connectivity, and Competition* (Cambria 2018). In 2019 and 2020, he served as the senior research director and lead author for the US Cyberspace Solarium Commission. He is a frequent contributor to *War on the Rocks*.

Audun Jøsang is a professor and head of the Research Group on Digital Security at the University of Oslo. Previously, he was an associate professor at QUT and a research leader for information security at DSTC in Australia. He has also worked in the telecommunications industry for Alcatel Telecom in Belgium and Telenor in Norway. Prof. Jøsang has a PhD in information security from NTNU, where he also worked as an associate professor. He has two master's degrees – one in information security from Royal Holloway College, University of London, and the other in telecommunications from NTH. He is a CISSP and CISM with broad expertise and experience in information security.

Monica Kaminska is a postdoctoral researcher at The Hague Program for Cyber Norms at Leiden University – Institute of Security and Global Affairs and a PhD candidate in Cyber Security at the University of Oxford. She is also a trustee of the European Cyber Conflict Research Initiative (ECCRI) and has held research positions at the Centre for Technology and Global Affairs and the Computational Propaganda Project, both at the University of Oxford. Kaminska's research examines international cyber conflict, particularly state responses to hostile cyber operations.

Johannes Klick is the CEO of Alpha Strike Labs and a passionate researcher and IT security expert. His research focuses on global internet scanning for threat analysis and external attack surface detection, as well as identifying the cyber infrastructures of organisations and companies. He has given presentations at internationally recognised conferences, such as Chaos Communication Camp, PHDays Russia, and Blackhat USA.

Commander PD Dr rer. nat. Dr habil. **Robert Koch** is a General Staff Officer of the German Federal Armed Forces. He joined the Navy in 1998 and studied computer science at the Universität der Bundeswehr. After operational and technical training, he served as a Weapon Engineering Officer onboard German frigates. After the National General/Admiral Staff Officer Course at the Bundeswehr Command and Staff College, Dr Koch had assignments as Action Officer Cyber at the Bundeswehr Communication and Information Systems Command and as Head of Department Penetration Testing at the Bundeswehr Cyber Security Center. Currently, Dr Koch is Desk Officer Cyber Policy at the Federal Ministry of Defence. He holds a Diploma in Computer Science

and a Master's in Military Leadership and International Security; he received his PhD in 2011, his habilitation in 2017 and his Venia Legendi in 2018. Now, he is lecturer in computer science at the Universität der Bundeswehr and the University of Bonn with course offerings ranging from introductory cyber security for non-technical people to in-depth technical IT security topics. His main areas of research are attack and tamper detection, anonymity in cyberspace, and conducting and optimising security analysis.

Csaba Krasznay is an associate professor at the University of Public Service with cybersecurity being his field of research. Currently, he is also the head of the university's Institute of Cybersecurity. Besides his activities in higher education, he is also present on the market. He obtained CISA certification in 2005, CISM and CISSP in 2006, CEH in 2008, ISO 27001 Lead Auditor in 2012 and CSSLP in 2015. He is a board member of the Magyary Zoltán E-government Association and of KIBEV – the Hungarian Voluntary Cyberdefence Collaboration, member of ISACA Budapest Chapter and the Hungarian Association for Electronic Signature. In 2011, he was voted the Security Expert of the Year.

Neal Kushwaha is a recipient of the 2019 Royal Humane Silver Medal of Bravery and 2020 commendation from the Governor General of Canada. He is an international bilingual speaker, a guest lecturer at Stellenbosch University, and a motivational speaker. He is also the founder of a Canadian consulting company, IMPENDO Inc., specifically serving public sector clients. His research and consulting falls within the cross-section of cyberspace, security, and law. Neal has published 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 over 28 years of climbing experience and climbs under the banner of Big Climbs, including Everest.

Artūrs Lavrenovs is a security researcher at the NATO CCDCOE focusing on the web and network technologies while teaching security courses, performing applied and academic research, and contributing to cyber exercises. His current research interests include internet measurement, connected device classification, and measuring distributed denial-of-service attack potential. Mr Lavrenovs has taught web and technology courses at the University of Latvia, where he is currently a PhD candidate.

Vincent Lenders (PhD) is the Director of the Cyber-Defence Campus at the Federal Office of Defence Procurement of Switzerland (armasuisse). Dr Lenders completed his PhD (2006) and MSc (2001) in electrical engineering and information technology at ETH Zurich and has been a postdoctoral researcher at Princeton University (2007). Vincent Lenders' interests lay at the intersection between cyber security, artificial

intelligence, networking, and crowdsourcing. His contributions over the last 15 years in these areas are documented in more than 130 internationally peer-reviewed research publications and have been implemented in various systems at the Swiss Armed Forces. He is also the co-founder and acts as a board member of the OpenSky Network and Electrosense associations.

Martin C. Libicki (PhD, UC Berkeley 1978) holds the Keyser Chair of Cybersecurity Studies at the United States 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 twelve years at the National Defense University, three years on the Navy Staff (logistics) and three years for the US GAO.

Marilia Maciel is a Digital Policy Senior Researcher at DiploFoundation. Previously, she was a researcher and coordinator in the Center for Technology and Society at the Getulio Vargas Foundation (CTS/FGV) in Rio de Janeiro. From 2017 to 2019, she served as deputy chair of the Research Advisory Group on Internet Governance in the Global Commission on the Stability of Cyberspace (GCSC). She has more than 15 years of experience in the area of digital policy, conducting research and writing policy briefs, legal opinions and academic papers, as well as developing and delivering capacity development. Her PhD research focuses on the securitisation of trade policy.

Ivan Martinovic is a professor at the Department of Computer Science, University of Oxford. Previously, he was a postdoctoral researcher at the Security Research Lab, UC Berkeley, and at the Secure Computing and Networking Center, UC Irvine. From 2009 until 2011 he held a Carl-Zeiss Foundation Fellowship and was an associate lecturer at TU Kaiserslautern, Germany. He obtained his PhD from TU Kaiserslautern under the supervision of Prof. Jens B. Schmitt and his MSc from TU Darmstadt, Germany.

Vasileios Mavroeidis is a research scientist and educator in cyber security at the University of Oslo, specialising in security automation and cyber threat intelligence generation, representation and sharing. Mr Mavroeidis has been affiliated with multiple national and European cyber security research and innovation projects in various roles and has been actively involved in cyber security standardisation.

Roland Meier is a fifth-year PhD student at the Department of Electrical Engineering and Information Technology at ETH Zürich. His research focuses on the security of

computer networks. In particular, he works on solutions that leverage recent advances in network programmability to make networks able to detect and mitigate attacks in the data plane and to provide more security and privacy. Mr Meier received his master's degree in electrical engineering and information technology from ETH Zürich in 2015.

James Pavur is a DPhil student and Rhodes Scholar at Oxford University's Department of Computer Science. His research focuses on satellite cyber security and ways to secure modern space missions from emerging threats. His work on satellite broadband security has been presented at leading technical venues, such as Black Hat and IEEE S&P, as well as receiving coverage in the popular press. Prior to attending Oxford, he studied science, technology and international affairs at Georgetown University's School of Foreign Service. In his free time, Mr Pavur is an avid competitive hacker, participating in hackathons and capture-the-flag competitions.

Nikola Pijović (PhD) is a Cyber Security Cooperative Research Centre Postdoctoral Research Fellow at the School of Politics and International Relations and the Law School, University of Adelaide. Dr Pijović is an expert on foreign policy and international relations, and his current research focuses on international cyber security policy in the Five Eyes.

Vladimir Radunović is a lecturer and director of educational programmes in the areas of internet governance, cyber security and e-diplomacy at the Diplo Foundation, an international non-governmental institution. He was a member of the Advisory Board of the Global Forum on Cyber Expertise and a member of the UN Multistakeholder Advisory Group for the Internet Governance Forum. His professional focus is on digital policies and diplomacy, in particular cyber security, cyber diplomacy, education and capacity-building for institutions and internet decision-makers. He graduated from the Faculty of Electrical Engineering in Belgrade and has a master's degree in contemporary diplomacy from the University of Malta. He is a PhD candidate in cyber security.

Lt. Col. **Anastasia Roberts** has been a legal officer in the British Army for 16 years. She served in both Afghanistan and Iraq. She is currently posted to the Office of Legal Affairs at SHAPE, as part of the Operational Law Branch. Her particular area of specialist expertise is legal support to Cyber, Intelligence and Information Operations.

James Shires (PhD) is an assistant professor at the Institute for Security and Global Affairs, University of Leiden, a fellow with the Cyber Statecraft Initiative at the Atlantic Council, and an associate fellow with The Hague Program for Cyber Norms. He holds a PhD in International Relations from the University of Oxford, an MSc from

Birkbeck College, University of London, and a BA from the University of Cambridge. He has written extensively on cybersecurity and international politics, most recently on disinformation and hack-and-leak operations. He is the author of *The Politics of Cybersecurity in the Middle East* (Hurst, 2021).

Tassilo Singer (PhD) is an IT Consultant. Dr Singer advises government institutions and international organisations on cyber security, cyber operations and AI. He has been a research associate and lecturer at the University of Passau (2012–2013, 2015–2017) and at the European University Viadrina (2013–2015), where he obtained his PhD on the dehumanisation of warfare and the necessity of human control. He has been guest lecturer at universities across Europe and taught international law in the context of cyber operations for government agencies. His research interests cover modern weapon technologies, AI solutions, data architecture and information security and related policies.

Martin Strohmeier is a Senior Scientist at the Swiss Cyber-Defence Campus. The main focus of his work has been the design, implementation, and analysis of security in cyber-physical systems, specifically in critical infrastructures. During his PhD at Oxford, he has extensively analysed the security and privacy of wireless aviation technologies. Subsequently, he has extended his interests towards areas of open-source intelligence, privacy issues in aviation and satellite environments, and adversarial machine learning. Martin is also a co-founder and board member of the aviation research network OpenSky. Before coming to Oxford in 2012, he received his MSc degree from TU Kaiserslautern and joined Lancaster University's InfoLab21 and Lufthansa AG as a visiting researcher. His work on aviation security has received several awards from the aviation and computer security communities.

Col. **Paul N. Sullivan** is a retired military officer (commander) having served 37 years in the military. He has worked with the Army Research Laboratory's Collaborative Technology Alliance, Collaborative Research Alliance and the International Technology Alliance. He has served at the Defense Research Projects Agency and is the author of several DOD directives and courses on military intelligence operations. Col. Sullivan has co-authored over 25 internationally published papers and presentations and has been involved in every major US military operation since the 1970s. He is former chairman of the S&T Working Group for the Director of Central Intelligence.

Brandon Valeriano (PhD Vanderbilt University) is the Donald Bren Chair of Military Innovation at the Marine Corps University. He also serves as senior adviser to the Cyber Solarium Commission and a senior fellow at the Cato Institute. Dr Valeriano has published six books and dozens of articles, including *Cyber War versus Cyber Reality* (2015) and *Cyber Strategy* (2018), both with Oxford University Press.

Bobby Vedral is an investor and founder of MacroEagle, a macro-political newsletter widely read in the financial community. He has over 20 years of experience in investment banking, most recently as a partner at Goldman Sachs. He holds an MA in international relations from the London School of Economics and a BA in Business Administration from ESB Reutlingen/ICADE. He is currently a PhD candidate in Modern War Studies at Buckingham University.

Adrian Venables (PhD) served in the Royal Navy for 24 years as a Communications, Warfare and Intelligence officer. Since leaving the Service, he has published a series of journal articles and research papers on the cyber threat landscape and its use by state and non-state actors for espionage, sabotage and subversion. Joining TalTech (Tallinn University of Technology) in Estonia as a senior researcher in 2018, he specialises in cyber strategy and its role in influence operations. Dr Venables retains his military links by serving as a Commander in the Royal Naval Reserve supporting UK cyber resilience activities in the Baltic region.

Skanda Vivek (PhD) has been an assistant professor of physics at Georgia Gwinnett College since 2019. Dr Vivek's research involves interdisciplinary approaches to revealing cyber-physical vulnerabilities across complex networks and building cyber-resilient societies. Prior to his current position, he was a postdoctoral research fellow at GeorgiaTech, where he applied statistical physics to unveil hidden collective risks in the event of cyber attacks on connected vehicles. He received his PhD in physics from Emory University in 2016. Dr Vivek's research has been published in multiple scientific journals including Nature, PNAS, PRE as well as broadcast widely by outlets such as the BBC and Forbes.

Bruce Watson (PhD) is chief scientist at IP Blox and adviser at Impendo, providing pattern recognition, algorithmic, and AI solutions for cybersecurity and complex-event recognition. Dr Watson also consults to public sector organisations on AI and cybersecurity, while being involved in startups and serving as research professor at Stellenbosch University (Centre for Artificial Intelligence Research and School for Data Science and Computational Thinking). Watson earned his first PhD in computing science and engineering from Eindhoven, after studying cryptography, graph algorithmics, and computer science in Waterloo, Canada. He later returned to Eindhoven as chair of Software Construction. Dr Watson's second PhD is from the University of Pretoria. Parallel to his academic career, he has 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, most recently moderating the Quantum Computing panel in 2018 and the Artificial Intelligence panels in 2019 and 2020.