‘Responsibility to detect?’: autonomous threat detection and its implications for due diligence in cyberspace

Sukumar, A.M.; Jancarkova, T.; Visky, G.; Winther, I.

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

'Responsibility to Detect?': Autonomous Threat Detection and its Implications for Due Diligence in Cyberspace

Arun Mohan Sukumar
PhD Candidate
The Fletcher School of Law and Diplomacy
Tufts University
Medford, MA, United States
arun.sukumar@tufts.edu

Abstract: Private and public organizations have long relied on intrusion detection systems to alert them of malicious activity in their digital networks. These systems were designed to detect threat signatures in static networks or infer anomalous activity based on their security 'logs'. They are, however, of limited use to detect threats across heterogeneous, modern-day networks, where computing resources are distributed across cloud or routing services. Recent advancements in machine learning (ML) have led to the development of autonomous threat detection (ATD) applications that monitor, evaluate, and respond to malicious activity with minimal human intervention. The use of 'intelligent' and programmable algorithms for ATD will reduce incident response times and enhance the capacity of states to detect threats originating from any layer of their territorial information and communications technologies (ICT) infrastructure. This paper argues that ATD technologies will influence the evolution of a due diligence rule for cyberspace by raising the standard of care owed by states to prevent their networks from being used for malicious, transboundary ICT activities. This paper comprises five sections. Section 1 introduces the paper and its central argument. Section 2 outlines broad trends and operational factors pushing public and private entities towards the adoption of ATD. Section 3 offers an overview of a typical ATD application. Section 4 analyses the impact of ATD on the due diligence obligations of states. Section 5 presents the paper's conclusions.

Keywords: *due diligence, autonomous systems, international law, threat detection, machine learning*

1. INTRODUCTION

The use of ATD technologies – applications capable of identifying, analysing, and in some cases, even responding to malicious activity in digital networks with minimal or no human supervision – has been contemplated for nearly three decades.[1] Today, with advancements in computing, it is possible to implement at scale ATD that relies on artificial intelligence (AI) / machine learning (ML) models. If computing advancements have made the widespread deployment of ATD possible, the pervasive digitalization of services and 'things' has made autonomous detection somewhat necessary. The topology of the modern digital network is extremely diverse, consisting of software and hardware whose ownership or management is often shared across vendors and geographies. In particular, the shift towards 'work-from-home', precipitated overnight by the COVID-19 pandemic, has resulted in businesses relying on cloud and network services that are scattered regionally and even globally. The terms 'network operator' and 'sysadmin' (short for 'system administrator') – the individual or organization responsible, inter alia, for managing the security of their enterprise's virtual infrastructure – are themselves misnomers in the contemporary era, given that they have limited visibility over vulnerabilities or attack vectors across network components. The traditional notion of a 'security perimeter', understood as the outer limits of a spatially and physically bound intranet, has suddenly become outdated.[2] Yet, threat detection tools have been slow to respond to this shift. Repetto et al. argue that cyber security applications are currently deployed in 'vertical silos' across a network, protecting the cloud service, applications, devices, enterprise architecture, etc.[3] This is not a viable solution to protect heterogeneous environments and often leads to malicious actors seeking out the path of least resistance in a network. ATD applications autonomously interface with various components of a digital network, drawing relevant and contextual information from those components to learn and detect potentially malicious threats. Consequently, these applications provide human operators with greater visibility over their fragmented network and obviate the cumbersome, manual monitoring of threats across various network components.

The roll-out of ATD in the public and private sector will shape not only cyber security practices but also the application of international law to state behaviour in cyberspace. This paper argues that the ability of ATD applications to detect and notify network operators of malicious activity will raise the standard of care owed by states to address significant transboundary harm in cyberspace. In other words, the adoption of ATD

1 See generally, H Debar, M Becker, and D Siboni, 'A Neural Network Component for an Intrusion Detection System' in *Proceedings of the 1992 IEEE Computer Society Symposium on Research in Security and Privacy*, 1992, 240–250; Jeremy Frank, 'Artificial Intelligence and Intrusion Detection: Current and Future Directions', in *Proceedings of the 17th National Computer Security Conference*, 1994, 1–12.

2 R. Rapuzzi and M. Repetto, 'Building Situational Awareness for Network Threats in Fog/Edge Computing: Emerging Paradigms beyond the Security Perimeter Model', *Future Generation Computer Systems* 85 (August 1, 2018): 235.

3 Matteo Repetto et al., 'An Autonomous Cybersecurity Framework for Next-Generation Digital Service Chains' (2021) 4 *Journal of Network and Systems Management* 29, 36.

will decisively influence the cyber due diligence principle. This paper comprises five sections. Section 2 outlines broad trends and operational factors pushing public and private entities towards adopting ATD to monitor their networks. Section 3 offers an overview of a typical ATD framework. Section 4 analyses the impact of ATD on the due diligence obligations of states. Section 5 presents the paper's conclusions.

2. THREAT DETECTION IN FRAGMENTED NETWORKS

The 'virtualization' of computing functions has lent incredible complexity to the modern-day digital network. As Valenza et al. note, there is no longer a linear connection that can be drawn between the digital application and the end-user device:[4] if its data is processed in cloud or edge servers, its networking functions are also outsourced for reasons of efficiency and costs. The fragmentation of the digital network makes it difficult for any one operator to monitor the entire network's security. This section reviews some of the factors identified in the technical literature that make conventional threat detection challenging in the contemporary era.

A. Multi-tenancy

Multi-tenancy refers to the sharing of software and hardware resources by a group of entities. These resources usually take the form of servers or databases used for storing or processing data. Cloud computing is the most common example of a multi-tenant framework where an umbrella vendor, such as Amazon Web Services (AWS) or Microsoft Azure, hosts several clients on its servers. In a multi-tenant environment, the 'tenants' manage their own end-point protocols, potentially creating multiple vectors of vulnerability for the whole infrastructure.[5] Tenants are also heavily dependent on the resources of the host, creating a central point of cyber security failure.[6] The challenges in securing multi-tenant environments are not limited to commercial networks but also extend to industrial control systems and critical infrastructure (CI). Indeed, the dependency of CI providers on cloud-based services has elicited stringent policy measures from states. For example, in December 2021, Australia passed legislation allowing government agencies to commandeer the resources of 'critical infrastructure assets' – including private cloud operators – to respond to a serious cyber attack in

4 Fulvio Valenza, Matteo Repetto and Stavros Shiaeles, 'Guest Editorial: Special Issue on Novel Cyber-Security Paradigms for Software-Defined and Virtualized Systems' *Computer Networks* 193 (July 5, 2021): 108126.

5 Repetto et al. (n 3) 37.

6 Wayne J Brown, Vince Anderson and Qing Tan, 'Multitenancy – Security Risks and Countermeasures' in *2012 15th International Conference on Network-Based Information Systems*, 2012, 7–13; see generally, 'SolarWinds Breach Exposes Hybrid Multicloud Security Weaknesses' (VentureBeat, 16 May 2021) <https://venturebeat.com/2021/05/16/solarwinds-breach-exposes-hybrid-multi-cloud-security-weaknesses/> accessed 10 November 2021.

exigent circumstances.[7] Such *post facto* measures, however, still do not address the challenge of timely identification of cyber attacks on multi-tenant environments.

In addition to outsourcing data storage and processing, public and private entities also delegate networking functions to third parties. This delegation, called Network Functions Virtualization (NFV), involves the handing over of packet-switching and routing functions, and even network-associated services such as firewall security, to an external vendor, such as Cloudflare, Akamai, or VMWare. NFV allows small organizations to conserve costs associated with purchasing and maintaining networking infrastructure, but the diversity of hardware components also makes it difficult to 'isolate and contain malware' within their networks.[8] As Firoozjaei et al. note, NFV exposes networks not only to infrastructure-based threats but also to targeted end-user threats because network services such as firewalls or secure sockets layer (SSL) gateways give NFV providers 'complete dominance over the user's information'.[9] Internet of Things (IoT) systems in particular have come to depend on NFV, given the limited computing power of IoT devices and the requirement for low latency of data traffic in some cases.[10]

Multi-tenant services have limited incentives to guarantee the security of their clients. As Schneier and Herr note, 'security is largely an externality' for cloud vendors because the cost of cyber attacks is borne by users and client organizations.[11] With CI, those costs are often borne by governments. Consequently, cyber attacks in multi-tenant environments are likely to persist, making timely identification all the more relevant.

B. Widespread Use of Application Programming Interfaces

Partly on account of the rise of multi-tenancy, but owing primarily to the explosive growth of the platform economy and social media, the internet is witnessing unprecedented 'API-fication'. Application programming interfaces (APIs) are lines of code that allow software to communicate with each other. They are digital railroads, built either by governments or private actors, that allow third parties to retrieve user data, integrate with multisided platforms, and in the case of NFV, communicate with

7 'Government Assistance' (Australian Government Department of Home Affairs) <www.homeaffairs.gov.au/about-us/our-portfolios/national-security/security-coordination/security-of-critical-infrastructure-act-2018-amendments/government-assistance> accessed 9 December 2021; 'Security Legislation Amendment (Critical Infrastructure) Act 2021', No. 124, 2021 (The Parliament of the Commonwealth of Australia), 63-74, <https://www.legislation.gov.au/Details/C2021A00124> accessed 10 November 2021.

8 'What Is Network Functions Virtualization (NFV)? | VMware Glossary' (VMware) <www.vmware.com/topics/glossary/content/network-functions-virtualization-nfv.html> accessed 9 January 2022.

9 Mahdi Daghmehchi Firoozjaei et al., 'Security Challenges with Network Functions Virtualization' (2017) 67 *Future Generation Computer Systems*, 315, 320.

10 See Nikos Bizanis and Fernando A Kuipers, 'SDN and Virtualization Solutions for the Internet of Things: A Survey' (2016) 4 *IEEE Access* 5591–5606.

11 Bruce Schneier and Trey Herr, 'Russia's Hacking Success Shows How Vulnerable the Cloud Is' (*Foreign Policy*, 24 May 2021) <https://foreignpolicy.com/2021/05/24/cybersecurity-cyberattack-russia-hackers-cloud-sunburst-microsoft-office-365-data-leak/> accessed 10 November 2021

applications as well as routing infrastructure in order to direct network traffic.[12] APIs play an important role in ensuring the interoperability of digital networks and seamless delivery of digital services. API 'calls' make up 83% of online traffic today.[13] With the proliferation of APIs, however, have emerged attendant risks. While acting as the internet's connective tissue, APIs also considerably expand the cyber attack surface.[14] Specifically, APIs expose networks to data breaches (the most common API security incidents[15]), person-in-the-middle attacks, and Distributed Denial of Service (DDoS) attacks that disrupt the availability of services, among others.[16]

Despite these concerns, API security has largely been sidestepped in favour of ease of adoption.[17] Weak authentication mechanisms, data leakage, and poor auditing of APIs have already led to major cyber security incidents and pose a challenge for businesses and policymakers alike.[18] The diversity and differential security policies of APIs, especially in fragmented networks, make threat detection extremely difficult.

C. Uneven Cyber Security Policy Landscape

Despite vulnerabilities posed by multi-tenant frameworks and the widespread use of APIs, most states have traditionally equated cyber security with data protection at the 'last mile'.[19] National regulatory instruments often focus exclusively on the

12 Truman Boyes et al., 'Accelerating NFV Delivery with OpenStack: Global Telecoms Align Around Open Source Networking Future' (OpenStack Foundation Report, 2016) <https://object-storage-ca-ymq-1.vexxhost.net/swift/v1/6e4619c416ff4bd19e1c087f27a43eea/www-assets-prod/marketing/OpenStack-NFV-Print.pdf> accessed 3 November 2021.

13 Akamai, 'State of the Internet / Security: Retail Attacks and API Traffic Report' (2019) <https://www.akamai.com/content/dam/site/it/documents/state-of-the-internet/state-of-the-internet-security-retail-attacks-and-api-traffic-report-2019.pdf> accessed 3 November 2021.

14 'Akamai: API: The Attack Surface That Connects Us All' (2021) 11 *Computer Fraud & Security* 4.

15 Brian Krebs, 'USPS Site Exposed Data on 60 Million Users – Krebs on Security' (*KrebsonSecurity* 21 November 2018) <https://krebsonsecurity.com/2018/11/usps-site-exposed-data-on-60-million-users/> accessed 3 November 2021; Dan Salmon, 'I Scraped Millions of Venmo Payments. Your Data Is at Risk' (*Wired*, 26 June 2019) <www.wired.com/story/i-scraped-millions-of-venmo-payments-your-data-is-at-risk/> accessed 3 November 2021; 'Rapid Growth of APIs Has Led to Security Risks for the Enterprise' (Cloudflare) <https://www.cloudflare.com/insights-api-proliferation/> accessed 9 January 2022.

16 Torsten George, 'The Next Big Cyber-Attack Vector: APIs' (*SecurityWeek*, 28 June 2018) <www.securityweek.com/next-big-cyber-attack-vector-apis> accessed 3 November 2021; 'API Attacks Are Both Underdetected and Underreported' (*Help Net Security*, 28 October 2021) <www.helpnetsecurity.com/2021/10/28/security-concerns-api/> accessed 3 November 2021.

17 See 'API Data Breaches in 2020' 9*CloudVector*, 23 December 2020) <www.cloudvector.com/api-data-breaches-in-2020/> accessed 5 November 2021.

18 Lindsey O'Donnell, 'Microsoft OAuth Flaw Opens Azure Accounts to Takeover' (*Threatpost*, 2 December 2019) <https://threatpost.com/microsoft-oauth-flaw-azure-takeover/150737/> accessed 9 November 2021]; 'Uber Disclosed on HackerOne: Sensitive User Information Disclosure at Bonjour.Uber.Com/Marketplace/_rpc via the "userUuid" Parameter' (HackerOne) <https://hackerone.com/reports/542340> accessed 9 January 2022; 'Amazon's Ring Neighbors App Exposed Users' Precise Locations and Home Addresses' (*TechCrunch*) <https://social.techcrunch.com/2021/01/14/ring-neighbors-exposed-locations-addresses/> accessed 9 January 2022; 'Information Leakage in AWS Resource-Based Policy APIs' (*Unit42*, 17 November 2020) <https://unit42.paloaltonetworks.com/aws-resource-based-policy-apis/> accessed 10 November 2021.

19 Jeff Kosseff, 'Defining Cybersecurity Law' (2018) 103 Iowa L. Rev. 985, 995.To be sure, this has changed in recent years. See generally, Agnes Kasper and Alexander Antonov, *Towards Conceptualizing EU Cybersecurity Law*, Discussion Paper / Zentrum Für Europäische Integrationsforschung, C 253 (Bonn: Zentrum für Europäische Integrationsforschung, Rheinische Friedrich-Wilhelms Universität, 2019) 26.

relationship between the end-user and their device or application, laying down broad guidelines on the types of personal and non-personal data that private and public entities can collect and store. Other network layers and infrastructure are often ignored. As a result, outside of data breaches, API development or the roles and responsibilities of cloud and NFV service providers have been poorly regulated in most jurisdictions. The lack of a clear regulatory framework on this issue makes vulnerabilities and incident reporting largely a factor of market practices, which are hardly uniform within and across states.

3. ATD: THE FUTURE OF THREAT DETECTION

Confronted thus by a fragmented network environment and uneven policy standards, states and private actors may currently pursue three options for whole-of-network threat detection.

Layered security: Organizations can depend on separate services to monitor and protect their network infrastructure, cloud-based resources, applications, and terminal devices. Given the difficulty and costs involved in integrating threat inputs from different sources, this approach is unlikely to be preferred by most network administrators.

Host-based security: Service providers such as AWS, Alibaba Cloud, and Cloudflare have begun offering services that monitor network traffic, perform authentication, and track API security.[20] Most of these services allow network administrators to 'remotely manage'[21] incidents from a centralized console.

Third-party security: Network operators can also rely on the services of a third party, which is usually a commercial entity (for example, Checkpoint, CrowdStrike, Mandiant). Third-party security offers flexibility to organizations that may want to avoid a lock-in of their threat detection capabilities with the host that provides multi-tenant services. Additionally, many cyber security vendors claim to offer multi-cloud threat detection, allowing clients to identify anomalous behaviour across various services.[22]

20 'Security at the Edge: Core Principles' (AWS, 24 September 2021) <https://d1.awsstatic.com/whitepapers/Security/security-at-the-edge.pdf> accessed 3 November 2021; 'Getting Started with Secure Access Service Edge: A Guide to Secure and Streamline Your Network Infrastructure' (Cloudflare, 22 October 2021) <www.cloudflare.com/static/52527ba193cc7ab0da6c23075d093ab3/Cloudflare_One_SASE_Whitepaper.pdf> accessed 3 November 2021; 'Alibaba Cloud Security Services' (Alibaba Cloud) <www.alibabacloud.com/product/security> accessed 9 January 2022.

21 See 'AWS IoT Device Management Features – AWS' (Amazon Web Services, Inc.) <https://aws.amazon.com/iot-device-management/features/> accessed 9 January 2022.

22 'Cloud Native Security – Security Automated Everywhere' (Checkpoint, 2021), <www.checkpoint.com/downloads/products/cloudguard-cloud-native-security-datasheet.pdf> accessed 3 November 2021.

In the wake of the COVID-19 pandemic, more organizations have sought 'full-stack observability'[23] over disparate components of their network. They may move towards host-based and third-party security offerings, as described above. These services have, in turn, begun implementing AI/ML models to perform intrusion detection.[24] AI/ML-based detection turns the problem of multi-tenancy into a solution. As with other aspects of digital networking, AI-based security functions are also increasingly offloaded to cloud or edge servers. With more computing resources available to process large volumes of traffic, it is today possible to train algorithms 'remotely' to detect threats and respond to them with low latency. This is especially useful in the case of IoT networks, which have increasingly been targets of DDoS attacks.[25]

However, the use of AI/ML models to detect cyber security threats has hitherto tended to fall into two categories: algorithms that can detect anomalies in 'static' topologies, that is, those networks where routing is predictable and where ports of entry and exit remain constant, or algorithms that learn to detect very specific malware in a network, whether it is based on unusual signatures or traffic patterns.[26] In dynamic multi-tenant environments, such applications are of limited use.[27] Nevertheless, newer applications of ATD leverage advancements in edge and cloud computing to obtain greater visibility over heterogeneous network environments. Such applications lean on a centralized architecture that performs whole-of-network monitoring, irrespective of the changing elements of its infrastructure or applications. The following paragraphs review the functioning of a typical ATD application.

In simple terms, ATD applications fetch information from various components of the network into 'clean rooms' that subsequently process such data to identify threats. ASTRID, a multistakeholder pilot project supported by the European Union, offers an example of such an ATD application.[28] ASTRID – which stands for AddreSsing ThReats for virtualIzeD services – provides a conceptual and technical framework to 'decouple' security functions from the overall functioning of individual network components. It does so by creating a 'centralized architecture' that collects 'security information, data, and events' from various network sources. This architecture

23 Erwan Paccard, 'Why Full-Stack Observability Is Critical for a Successful DevSecOps Approach' (*Computing*, 15 December 2021) <www.computing.co.uk/sponsored/4042095/stack-observability-critical-successful-devsecops-approach> accessed 12 November 2021.

24 'Endpoint Protection Software Explained' (Crowdstrike) <www.crowdstrike.com/cybersecurity-101/endpoint-security/endpoint-protection-software/> accessed 9 January 2022; 'Endpoint Protection Buyer's Guide 2020' (Checkpoint) <https://app.hushly.com/runtime/content/JGq2xOVJoBplBawJ> accessed 12 November 2021; 'Amazon Detective – AWS' <https://aws.amazon.com/detective/?c=sc&sec=srv> accessed 9 January 2022.

25 Liang Xiao et al., 'IoT Security Techniques Based on Machine Learning: How Do IoT Devices Use AI to Enhance Security?' 2018 35(5) *IEEE Signal Processing Magazine*, 41–49.

26 Maruthi Rohit Ayyagari et al., 'Intrusion Detection Techniques in Network Environment: A Systematic Review' (2021) 27(2) *Wireless Networks*, 1269.

27 Daniel Spiekermann and Jörg Keller, 'Unsupervised Packet-Based Anomaly Detection in Virtual Networks' (2021) 192 *Networks* 2.

28 'ASTRID Project: A Cybersecurity Framework for Virtualized Services' <https://www.astrid-project.eu/project.html> accessed 9 January 2022.

comprises a data plane, control plane, and management plane.[29] The data plane is the programmable component of the framework that collects and maintains security-related logs, events, and traffic metrics spanning the network. Most components of the network have event- and log-reporting capabilities built into their software, and the data plane relies on 'lightweight hooks', that is, APIs, to query and retrieve security information from their kernels or libraries. The control plane is a collection of ML algorithms that retrieve information from the data plane, and through it, obtains 'complete visibility' over the network. These algorithms evaluate the security information and identify threats based on anomalous behaviour. Finally, the management plane is the human-facing element of this architecture, which communicates threats in real time to network administrators and suggests remedial measures.

The operationalization of such an ATD architecture will depend on two technical factors. First, it requires the availability of adequate computing resources to perform 'clean room' functions. The data plane does not simply collect logs and events from network components but also dynamically adjusts the scope and frequency of reporting as necessitated by circumstances.[30] If potentially malicious behaviour is identified in one section of the network – for example, suspicious API calls or unusual router volumes – then the ATD application channels greater detection and remedial resources towards it. Similarly, if the network relies on a new cloud or NFV service, the ATD application may seek more data from it to learn its behaviour and train its algorithms. Such adjustments necessitate adequate computing resources. Second, ML-driven threat detection through a 'command-and-control' architecture requires interfacing with security functions of other network components (for example, firewalls, packet inspection tools, other analytics software, etc.). This is only possible if the various network services adopt common and interoperable interfaces allowing the data plane to ping and access relevant reporting information.

Both technical factors are close to realization today. As already noted, advancements in edge and cloud computing enhance the programmability of ATD applications. With respect to common security interfaces, there has been a notable parallel effort from within the technical community to develop interoperable standards that monitor heterogeneous networks. Since 2014, a Birds of a Feather (BoF) group in the Internet Engineering Task Force (IETF) has sought to promote discussion on a common 'Interface 2 Network Security Functions' (I2NSF).[31] This group, which includes volunteers from prominent global technology companies, has sought 'a standardized interface to control and monitor the rule sets that network security functions [NSFs]

29 ASTRID Consortium, 'D1.2 – ASTRID Architecture' 31–33 <https://cyberwatching.eu/sites/default/files/D1.2%20-%20ASTRID%20architecture.pdf> accessed 12 November 2021.

30 R Bolla, A Carrega, and M Repetto, 'An Abstraction Layer for Cybersecurity Context' in *2019 International Conference on Computing, Networking and Communications (ICNC)* (2019) 215.

31 'RFC 8192 Interface to Network Security Functions (I2NSF): Problem Statement and Use Cases' (IETF Datatracker) <https://datatracker.ietf.org/doc/rfc8192/> accessed 9 January 2022.

use to treat packets traversing through these NSFs'.[32] In other words, I2NSF aims to create 'vendor-agnostic' protocols that allow for a seamless flow of threats-related information – whether through centralized or distributed architecture – among various network components by accessing their security functions and capabilities. The BoF group has specifically emphasized the need for an interface sensitive to periodic updates of security policies or configurations by stand-alone network services, which is crucial to an 'autonomous security system'.[33] Indeed, given the synergies between both endeavours, the ASTRID project highlights in detail the characteristics of the I2NSF proposal.[34] Whether or not this specific IETF initiative succeeds,[35] it is only reasonable to conclude similar efforts – including those by the private sector[36] – will mushroom in the coming years.

ATD frameworks such as ASTRID enhance the visibility of system administrators over their networks and, through their programmability, also offer the system administrators greater control when addressing threats and vulnerabilities unique to their organizations. Beyond market consequences, the policy impact of ATD is equally notable. For instance, the 'command-and-control' model of ATD applications allows states to detect and respond to cyber security threats to publicly owned CI, even if such CI relies on private cloud/NFV services or even if those services are located in another country. ATD could also provide states with accurate and instantaneous knowledge of transboundary malicious activity emanating from or transiting through their territory. The following section addresses this possibility in greater detail.

How exactly could states rely on ATD frameworks? Some states may develop a 'plug-and-play' ATD application, using its technical framework to enforce cyber security policies on monitoring networks for harmful activity, and require all private and public operators based in its territory to adopt it. Other states could develop APIs for their Computer Security Incident Response Teams (CSIRTs) that interface with private ATD applications. As a result, CSIRTs would be automatically notified whenever potentially malicious threats are detected by those applications.

32 S Hares et al., 'Interface to Network Security Functions (I2NSF): Problem Statement and Use Cases' (RFC Editor, July 2017) 7 <www.rfc-editor.org/rfc/pdfrfc/rfc8192.txt.pdf> accessed 12 November 2021

33 'Re: [I2nsf] I2NSF Re-Chartering Text' <https://mailarchive.ietf.org/arch/msg/i2nsf/rn1F7BSqqEzI15ApV2c0UHjmbz8/> accessed 9 January 2022.

34 ASTRID Consortium (n 29) 68–72.

35 For an overview of the Internet Engineering Steering Group's competing views on the proposal, see 'Ballot for Draft IETF RFC (Interface to Network Security Functions (I2NSF): Problem Statement and Use Cases)' (IETF Datatracker) https://datatracker.ietf.org/doc/rfc8192/ballot/> accessed 9 January 2022.

36 Jordan Novet, 'Amazon's Outage and HashiCorp's IPO Point to a Future with Multiple Clouds' (CNBC, 12 December 12, 2021), https:// <www.cnbc.com/2021/12/12/aws-outage-and-hashicorp-ipo-point-to-a-multicloud-future.html> accessed 9 November 2021.

4. ATD AND CYBER DUE DILIGENCE

By enhancing their ability to detect harmful cyber activity, ATD applications could also influence the scope of duties states have with respect to preventing their territory from being used to launch or relay cyber operations targeting another state. The due diligence principle in international law, as enunciated by the International Court of Justice in *The Corfu Channel Case*, refers to the obligation of a state 'not to allow knowingly its territory to be used for acts contrary to the rights of other States.'[37] The due diligence principle requires that states take all reasonable steps necessary to prevent and mitigate activities on their territory that could cause 'significant transboundary harm'.[38] The standard of care owed by states to prevent and stop harmful transboundary activities is proportionate to the risk involved in such activities. As this formulation suggests, the due diligence principle essentializes an obligation of conduct, and not an obligation of result, that is, one determined by the outcome of a state's efforts to prevent transboundary harm.[39] Nevertheless, two important considerations attach themselves to any domain-specific due diligence rule. The first, as recognized by the arbitration tribunal in the *Alabama* case,[40] is that the standard of care owed by states is not the same as that which they 'ordinary employ in their domestic concerns'.[41] The requirement of due diligence stems from the 'international duties' of states, and as such, it is not enough to treat activities causing transboundary harm in the same manner as those whose effects are territorial. Second, the International Law Commission (ILC) has noted that the due diligence principle also requires states to 'keep abreast of technological changes'.[42] The ILC's commentary to the 2001 Draft Articles on Prevention of Transboundary Harm from Hazardous Activities emphasizes perceptions of 'reasonable' or 'appropriate' measures to prevent transboundary harm may evolve because of advancements in science and technology.[43]

Both elements of the due diligence principle are relevant in the context of cyber security. A cyber-specific due diligence rule, if one exists, would impose on states an obligation to monitor and prevent cyber operations that cause significant

37 *Corfu Channel Case (United Kingdom v Albania)* (Judgment of 9 April) [1949] ICJ Rep 22.

38 International Law Commission, *Draft Articles on Prevention of Transboundary Harm from Hazardous Activities, with Commentaries*, Article 3, (2001) UN Doc. A/56/10.

39 Antal Berkes, 'The Standard of "Due Diligence" as a Result of Interchange between the Law of Armed Conflict and General International Law' (2018) 23(3) *Journal of Conflict and Security Law* 433–460; Timo Koivurova, 'Due Diligence' (last updated February 2010), in A Peters and R Wolfrum (eds), *The Max Planck Encyclopedia of Public International Law* (Oxford University Press 2008–), < https://opil.ouplaw.com/view/10.1093/law:epil/9780199231690/law-9780199231690-e1034?rskey=Nvr0gA&result=1&prd=MPIL> accessed 3 March 2022; cf Antonio Coco and Talita de Souza Dias, '"Cyber Due Diligence": A Patchwork of Protective Obligations in International Law' (2021) 32(3) *European Journal of International Law* 773.

40 Alabama Claims Arbitration (1872) 1 Moore Intl Arbitrations 495.

41 Richard Mackenzie-Gray Scott, 'Due Diligence as a Secondary Rule of General International Law' (2021) 34(2) *Leiden Journal of International Law* 343, 351.

42 International Law Commission, *Draft Articles on Prevention of Transboundary Harm from Hazardous Activities, with Commentaries*, 154 (2001) UN Doc. A/56/10.

43 ibid.

transboundary harm. This duty of care would extend beyond measures to address domestic cyber crime and include steps taken specifically to address transboundary ICT activities. Following the commentary to the ILC Draft Articles, a 'cyber' due diligence rule would also require that states progressively adopt new technologies to monitor and mitigate harmful transboundary activity on their networks. Needless to say, these considerations place strong positive obligations upon states to prevent their territory from being used for harmful cross-border activities. Partly because of the difficulty in implementing those obligations, states do not all agree (as of March 2022) on the existence of a due diligence rule for cyberspace.[44] Even those legal scholars who acknowledge the existence of a cyber due diligence principle 'recognize a more limited duty' than that applicable to other domains – namely, a duty only to stop cyber operations and not to 'prevent, or even monitor' them.[45]

Despite ambiguity as to the precise scope of a cyber due diligence principle, there appears to be growing consensus among states that they should not 'knowingly allow their territory to be used for internationally wrongful acts' using ICTs.[46] This unique formulation originally appeared in the consensus report of the 2015 UN Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (GGE), which can be considered the lodestar for non-binding and voluntary guidelines on state behaviour in cyberspace. The 2019–2021 UN GGE built on the 2015 report's norms and identified an expectation on states to take all 'appropriate, reasonably available, and feasible steps to… detect, investigate, and address' internationally wrongful acts emanating from or transiting through their territory, provided they are 'aware or notified in good faith' of such acts.[47] The norms articulated by the GGE are not binding, but they represent a clear expression of intent on the part of states to move towards a due diligence regime on cyberspace. This argument is further strengthened by the fact that the 'zero' and 'first' drafts of the 2019–2021 Open-Ended Working Group (OEWG) on ICT security declared states should 'ensure that their territory is not used by non-state actors acting on the instruction or under the control of a state to commit [internationally wrongful] acts'.[48] The language of the drafts reflected an attempt to align the due diligence requirement with that of state responsibility and restrict the scope of positive obligations only to those instances where cyber operations were

44 For an overview of national positions of key states on due diligence, see 'Due Diligence – International Cyber Law: Interactive Toolkit' (Cyber Law Toolkit) <https://cyberlaw.ccdcoe.org/wiki/Due_diligence> accessed 9 January 2022.

45 ibid.

46 'Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security', UN Doc. A/70/174 (2015), 8/17.

47 'Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security', UN Doc. A/76/135 (2021), 10/26.

48 Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security (2021) 'Draft substantive report (zero draft)', A/AC.290/2021/L.2, 6/18. Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security (2021) 'Substantive Report [FIRST DRAFT]', 14, <https://front.un-arm.org/wp-content/uploads/2021/03/210301-First-Draft.pdf> accessed 6 December 2021.

attributable to the state. This formulation was opposed on the same ground in the final negotiating session of OEWG in February 2021[49] and consequently moved to the Chair's Summary for lack of consensus, whereas the more expansive GGE formulation noted above was later adopted in June 2021. The rejection of the OEWG draft language arguably indicates a desire on the part of most states to develop a stand-alone, *sui generis* framework on cyber due diligence.

Critical to the realization and, indeed, effectiveness of a due diligence regime is conceptual clarity on what it means for a state to have 'knowledge' of harmful, transboundary cyber operations. Knowledge is not only a 'constitutive element' of due diligence,[50] but also an important technical consideration. However, there is currently no congruence between the legal and the technical thresholds of 'knowledge' required to operationalize the due diligence obligation. The GGE's guidance suggests states should be 'aware or notified in good faith' of such activity to trigger their due diligence obligations.[51] A state may legally be considered 'aware' of transboundary malicious activity based on its actual or constructive knowledge ('should have known') of the activity.[52] A state may have actual knowledge if it receives 'credible information that a harmful cyber operation is underway from its territory'.[53] In technical terms, however, the compromising of digital infrastructure – as with command-and-control servers and attack surfaces in the case of botnet attacks[54] – begins well before the attack commences. At this stage, both the motives of the attacker and their intended target are usually unclear.[55] In other words, the legal threshold of 'actual knowledge' of the originator or transit state is often too high to pursue a meaningful implementation of the due diligence principle. Once an attack has commenced, it may in fact be challenging for a state to terminate the activity without avoiding serious disruptions to its domestic digital infrastructure or services. In other words, a state with 'actual knowledge' of malicious activity could legitimately claim inability to exercise its due diligence obligations on the ground that the termination of such activity demands an unreasonable technical effort on its part. On the other hand, the determination of a state's 'constructive knowledge' about the transboundary effects of malicious activity solely on the basis of anomalous behaviour on its digital networks is technically difficult

49 See generally, 'Comments by Germany on the OEWG Zero Draft Report', 2, < https://front.un-arm.org/wp-content/uploads/2021/02/Germany-Written-Contribution-OEWG-Zero-Draft-Report_clean.pdf> accessed 4 January 2022 ; 'The Netherlands – Written Proposals to OEWG Zero Draft', 2, < https://front.un-arm.org/wp-content/uploads/2021/02/Netherlands-OEWG-informals-intervention-Feb-2021.pdf> accessed 4 January 2022.

50 Michael N Schmitt (ed), 'Due Diligence', in *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (2nd edn, Cambridge: Cambridge University Press 2017), Rule 6, 40.

51 *2019–2021 UN GGE Report*, n. 53, 10/26.

52 *Tallinn Manual 2.0* (n 50) Rule 6, 33.

53 ibid, Rule 6, 40.

54 See generally, Manos Antonakakis et al. (2017) 'Understanding the Mirai Botnet' in *26th USENIX Security Symposium (USENIX Security 17)*, 1094–1095, < https://www.usenix.org/system/files/conference/usenixsecurity17/sec17-antonakakis.pdf> accessed 12 November 2022.

55 See, Scott J Shackelford, Scott Russell and Andreas Kuehn, 'Unpacking the International Law on Cybersecurity Due Diligence: Lessons from the Public and Private Sectors' (2016) 17(1) *Chicago Journal of International Law* 20.

but legally plausible (based on intelligence inputs, past incidents, political relations with the affected state, etc.).[56] The standard of proof for a victim state to establish the territorial state's constructive knowledge, as Delerue notes, would be 'very high'.[57] Further, as this paper has highlighted in extensive detail, 'actual' or 'constructive' knowledge about malicious cyber activity on one element of a heterogeneous, multi-layered network cannot reasonably be tantamount to such knowledge of a targeted cyber operation. The lack of network visibility and control over various network components makes such determination difficult.

ATD does not offer a panacea to the problems highlighted above but will help bring the mitigation and prevention of transboundary digital harm operationally closer to expectations generated by the due diligence obligation in international law. First, ATD applications specifically allow for the detection of malicious cyber activity that causes transboundary harm. Second, and related, the adoption of ATD applications will help align legal and technical thresholds of 'actual' and 'constructive' knowledge of transboundary harmful activity. Third, ATD applications will operationally support an expansive cyber due diligence rule, the scope of which is not limited to halting harmful activity but also includes also a responsibility to prevent such activity. And fourth, ATD applications are useful not only in instances where harmful cyber operations originate in a state's territory but also where the operations transit through it.

Provided network security interface standards (such as I2NSF) are interoperable, ATD applications that run on them can detect unusual cyber activity in any virtualized component of the surveilled network, irrespective of the territory where such component is located. This will prove beneficial both to victim states and originator states. Victim states can pinpoint with greater precision the transboundary source of harmful cyber activity. Even if they cannot exercise control over the compromised infrastructure, they can promptly notify the originator state. ATD applications will not only guide originator states to malicious cyber activity and compromised infrastructure within their territory but could also indicate, through deep packet inspection,[58] the transboundary targets of such cyber operations. Given that they rely on ML models, ATD applications 'learn' and discern patterns from previous incidents involving transboundary harm and accordingly notify administrators when they detect similarly anomalous behaviour – at an early stage – on the surveilled network. In this manner, they ensure states have timely access to information necessitating their exercise of due diligence, and thus bring forward legal standards of 'actual' and 'constructive' knowledge to meet new technical realities. The predictive capability of ATD

56 Russell Buchan, 'Cyberspace, Non-State Actors and the Obligation to Prevent Transboundary Harm' (2016) 21(3) *Journal of Conflict and Security Law* 429, 441–442; *Tallinn Manual 2.0* (n 49) Rule 6, 41.

57 François Delerue, *Cyber Operations and International Law* (1st edn, Cambridge University Press 2020) 367.

58 A Carrega et al., 'Situational Awareness in Virtual Networks: The ASTRID Approach', in *2018 IEEE 7th International Conference on Cloud Networking (CloudNet)*, 3 https://ieeexplore.ieee.org/document/8549540> accessed 11 December 2021.

applications not only enhances the knowledge of states but could also help expand the scope of the cyber due diligence principle to include monitoring of networks and prevention of harmful transboundary activity. In other words, ATD applications make it technically feasible not only to monitor heterogeneous digital networks for anomalous behaviour but also to analyse in advance the precise nature of the threat and the harm it is likely to cause.

Finally, ATD applications could also help states detect malicious activity transiting through territorial networks.[59] Such detection is often quite challenging, given that traffic routing is usually automated and a factor of speed and availability of network resources, rather than of conscious choice to steer malicious activity through the servers of a particular country.[60] Knowledge of transiting traffic is then a determinant of superior intelligence and technical capacity as well as geopolitical attributes associated with a country's location. This perhaps explains why countries like the Netherlands and Singapore – major data transit points that both enthusiastically champion the norm on the 'public core of the internet'[61] and call on states to protect the availability and integrity of internet infrastructure that has transnational functionality – differ in their views on due diligence. The Netherlands claims a cyber due diligence rule already exists, whereas Singapore, which has to contend with the geopolitical volatility and cyber 'insecurity' of Southeast Asia,[62] has been more cautious, calling for greater clarity on the 'degree of knowledge' implicated by a due diligence rule.[63] ATD applications, which already interface with virtualization services agnostic of territorial location, would be well equipped to detect anomalous behaviour transiting through network infrastructure of those services.

59 For views of legal scholars on the responsibility of transit states, see *Tallinn Manual 2.0* (n 49), 33–34; Eric Talbot Jensen and Sean Watts, "Cyber Due Diligence," (2021) 73 *Oklahoma Law Review* 645, 696, 70; August Reinisch and Markus Beham, 'Mitigating Risks: Inter-State Due Diligence Obligations in Case of Harmful Cyber Incidents and Malicious Cyber Activity – Obligations of the Transit State' (2015) 58 *German Yearbook of International Law*, 101.

60 There may, of course, exist scenarios where a transit country's servers could be specifically targeted by state and non-state actors to thwart attribution.

61 Alexey Trepykhalin and Veni Markovski, 'Country Focus Report: The Netherlands and the "Public Core of the Internet"' (ICANN, 2021) <https://www.icann.org/en/system/files/files/ge-008-28may21-en.pdf> accessed 12 November 2021; 'Singapore's Written Comment on the Chair's Pre-Draft of the OEWG Report', <https://front.un-arm.org/wp-content/uploads/2020/04/singapore-written-comment-on-pre-draft-oewg-report.pdf> accessed 9 January 2022.

62 'Southeast Asia: Cyber Threat Landscape' (FireEye) <www.fireeye.com/offers/rpt-sea-threat-landscape.html> accessed 9 January 2022.

63 'Official Compendium of Voluntary National Contributions on the Subject of How International Law Applies to the Use of Information and Communications Technologies by States', UN Doc. A/76/136 (2021), 84/142 <https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf> accessed 3 March 2022.

5. CONCLUSION

With rapid advancements in computing, it appears probable that ATD across heterogeneous networks will soon be a reality. ATD applications can help realize and expand the due diligence obligation of states by facilitating the early detection and notification of transboundary digital harm. Not only would ATD applications offer states greater visibility over hardware and software elements of territorial networks, but they would also automate the reporting of potentially malicious activity. Critically, ATD applications will be capable of detecting new threats and attacks by learning from anomalous behaviour or signatures previously observed on known attack surfaces and vectors. By operationalizing the 'knowledge' component of cyber due diligence, ATD applications thus raise the standard of care owed by states for not only stopping but also preventing transboundary digital harm.

However, these applications too come with their own share of security concerns. Malicious attackers could corrupt the training data used by ATD algorithms, generating false positives or misleading results, which in turn depletes public trust in such technologies. The ATD framework reviewed in this paper, including the ASTRID project, also depends on APIs, whose security could be compromised with grave consequences for the integrity of the network as a whole. Finally, it is also possible that repressive or autocratic states may force the adoption of ATD applications with a view to engaging in deep inspection of private networks for surveillance, under the garb of security or performance of due diligence obligations.[64] These factors have to be carefully weighed against the widespread adoption of ATD. Additionally, such technologies may be inaccessible to developing countries for reasons of cost or export control restrictions, resulting in the uneven development and laggard adoption of cyber due diligence. Nonetheless, ATD responds to a pressing need to monitor diverse and multilayered networks. The deployment of intelligent algorithms to monitor and detect cyber security threats will not only optimize resources but also reduce response times for private and public entities. In the process, they may transform what it means for states to exercise diligence in the performance of their international obligations in cyberspace.

ACKNOWLEDGEMENTS

The title of the paper is inspired by that of the article 'The Responsibility to Inspect: Due Diligence in Cyberspace', which appeared on the website of the Observer Research Foundation in July 2016.

64 Delerue (n 57) 360–362.