itag Cyber Research Conference Ireland

Multidisciplinary Perspectives on Cybersecurity Research, Practice and Education

Proceedings of the 1st Cyber Research Conference Ireland, Galway, Ireland April 25, 2022

Michael Lang, Séamus Dowling and Ruth Lennon (editors)

Ollscoil na Gaillimhe University of Galway

Ollscoil Teicneolaíochta an Atlantaigh Atlantic Technological University

itag Innovation Technology AtlanTec Gateway

CYBER | IRELAND IRELAND'S CYBER SECURITY CLUSTER

Michael Lang, Séamus Dowling & Ruth Lennon (eds.) (2022)
Multidisciplinary Perspectives on Cybersecurity Research, Practice and Education: Proceedings of the 1st Cyber Research Conference Ireland, Galway, Ireland, April 25, 2022. Galway: University of Galway.

ISBN 978-1-911690-00-9 (eBook)

Published by:

University of Galway
James Hardiman Library
Newcastle
Galway
H91 TK33
Ireland

This book is available electronically in the University of Galway ARAN institutional research repository
http://aran.library.nuigalway.ie

OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

Contents

Preface

This volume contains the peer-reviewed proceedings of the inaugural Cyber Research Conference Ireland (CRCI) which was hosted by Atlantic Technological University (Galway) in partnership with the University of Galway and the itag Cyber Forum / Cyber Ireland West Chapter. The event took place on April 25, 2022 and attracted 80 delegates from academia, industry, and the public sector.

The purpose of the conference was to bring together, for the first time, the multidisciplinary community of researchers across the island of Ireland who are engaged with various aspects of cybersecurity, cybercrime and related areas. Researchers from the fields of computer science, business information systems, finance and risk management, psychology, criminology, law, human resources and other cognate disciplines were in attendance.

The initial idea for this national conference was proposed at an itag Cyber Forum meeting in October 2021 by Dr. Michael Lang of the School of Business & Economics, University of Galway, who observed that there has never been any previous event in Ireland that brought together cyber researchers from different disciplines on a unified platform. Séamus Dowling and Ruth Lennon of Atlantic Technological University supported the suggestion and came on board as conference co-chair and programme chair respectively.

The event was a tremendous success, providing a forum for young and established researchers to present their work and to build networks. It is intended that CRCI will become an annual event travelling around the regions of Ireland.

The 2022 conference was kindly sponsored by the itag Skillnet programme.

Pictured at the Cyber Research Conference Ireland 2022 were, from left: Eamonn Larkin, IBM (Chairperson of itag Cyber Forum), Ruth Lennon, Atlantic Technological University (Programme Chair); Séamus Dowling, Atlantic Technological University (Conference co-Chair); Michael Lang, University of Galway (Conference co-Chair); Caroline Cawley (CEO of itag).

Conference Officers and Committee

Conference Officers

Conference Co-Chair	Michael Lang	University of Galway
Conference Co-Chair	Séamus Dowling	Atlantic Technological University (Mayo)
Programme Chair	Ruth G. Lennon	Atlantic Technological University (Donegal)
Publicity Chair	Eamon Larkin	IBM, itag Board
Local Organising Chair	Caroline Cawley	CEO of itag

Steering Committee

Caroline Cawley	CEO of itag
Séamus Dowling	Atlantic Technological University (Mayo)
Michael Lang	University of Galway
Eamon Larkin	IBM, Leader of itag Cyber Forum
Ruth G. Lennon	Atlantic Technological University (Donegal)
Veronica Rogers	Atlantic Technological University (Sligo)

Programme Committee

Enda Barrett	University of Galway, Ireland
Lena Connolly	Zayed University, United Arab Emirates
Eoin Cullina	Atlantic Technological University (Galway), Ireland
Séamus Dowling	Atlantic Technological University (Mayo), Ireland
Catherine Friend	Institute of Art, Design & Technology, Ireland
Greg E. Gogolin	Ferris State University, USA
David Kreps	University of Galway, Ireland
Michael Lang	University of Galway, Ireland
Ruth G. Lennon	Atlantic Technological University (Donegal), Ireland
Doug Leith	Trinity College Dublin, Ireland
Seán McSweeney	Munster Technological University, Ireland
Stuart Millar	Rapid 7, Ireland
Liam Noonan	Technological University of the Shannon (Midlands Midwest), Ireland
Diarmuid Ó Briain	Institute of Techology, Carlow, Ireland
Oluwafemi Olukoya	Queen's University Belfast, UK
Maria Grazia Porcedda	Trinity College Dublin, Ireland
Andrew Power	Institute of Art, Design and Technology, Ireland
Veronica Rogers	Atlantic Technological University (Sligo), Ireland
Michael Schukat	University of Galway, Ireland
Barry Sheehan	University of Limerick, Ireland
Irina Tal	Dublin City University, Ireland
Christina Thorpe	Technological University Dublin, Ireland
Simon Woodworth	University College Cork, Ireland

Phantom or Menace: User Behaviors in Cybersecurity

Thomas Acton
Business Information Systems
J.E. Cairnes School of Business & Economics
NUI Galway
Galway, Ireland
thomas.acton@nuigalway.ie

Pratim Milton Datta
Management & Information Systems
Ambassador Crawford College of Business & Entrepreneurship
Kent State University
Ohio, USA
pdatta@kent.edu

Martin Hughes
Business Information Systems
J.E. Cairnes School of Business & Economics
NUI Galway
Galway, Ireland
martin.hughes@nuigalway.ie

***Abstract*—This paper proposes a specific process-based approach to a systematic literature review to scope extant research on user behavior in cybersecurity. Focusing on the socio- rather than technical aspects of cybersecurity, it aims to identify pertinent studies, identify a set of categories of behavioral concern, and propose a set of further studies to investigate these categories. Further, the study proposes to identify user-focused behavioral themes of particular concern to organizations and users, to provide insights on user behaviors that can impact effective cybersecurity.**

Keywords—cybersecurity, user behavior, attack, information systems

I. INTRODUCTION

Cybersecurity is rapidly assuming prominence and importance for organizations [1, 2], federal governance bodies [3], and nations [4], largely pivoted by the breakneck digital transformation during COVID-19. Recent security breaches have highlighted the negative impacts of weak cybersecurity on health systems, university education, global banking, and commercial data [5, 6]. In the face of growing cybersecurity attacks, research and industry have, however, often relegated cybersecurity as a predominantly technological solution. However, such relegation is essentially myopic.

In most cases, after significant investigation, breaches can be traced to the accidental or intended actions of the individual, with resultant access to technical systems and compromising security layers [7, 22]. Such actions include a person clicking on a malicious web link; exposing a password; not securing access to sensitive data or systems; failing to secure systems to prevent malware, viruses or trojan infection; poor adherence to security protocols in organizations; accidental data loss or transfer; use of infected software or visiting infected web pages; not maintaining up-to-date software patches; or other behaviors reliant on user-triggered execution. Other than zero-day flaws in technologies, slow fix rollout by technology providers, or the increasing capability and sophistication of lone or state-backed hackers, it is mainly user-behavior that facilitates breaches to otherwise secure systems [8-10].

The extant literature in the information systems and its main informing fields has to date assessed broad aspects of cybersecurity, including cyberattacks, cyber policies, mitigation, recovery, and other facets of preparedness, response, and emerging issues. On the socio-technical spectrum inherent to information systems, numerous studies have focused on technical aspects to cybersecurity, in the main outside the core information systems literature bases but within the software engineering body of knowledge generally; others have focused on social aspects and, although largely within information systems literatures, have taken an organizational or corporate perspective [11-13].

Recently, there have been some research efforts reviewing cybersecurity user behaviors (e.g. Simon [22]). However, [22] is epistemic and mainly covers the "partial" representational aspects of human factors rather than a comprehensive synthesis across abroad range of human behaviors related to cybersecurity. Therein lies a serious lacunae in a systematic examination of user behavior in cybersecurity breaches, especially when facing a deluge or attacks and a paucity of cybersecurity perils owing to user behaviors.

A lack of understanding of cybersecurity user behavior not only weakens our understanding of cybersecurity but also exposes our burgeoning digital economies to a slew of future malicious attacks. Cybersecurity efficacy is important, because it comprises a non-binary multivariate spectrum. Overall there is a need for more coherence, rooted in a core discipline with informing domains, with the robust inclusion of user-centred issues in cybersecurity and, in particular, the impact of user behaviors on cybersecurity efficacy [10].

II. MOTIVATION

Cybersecurity has received media attention largely in the context of cyberattacks. Corporate response to cyberattack has largely been reactive [14], seeking to patch weaknesses after they have been exposed and after an attack has occurred [15]. Governments, dealing with a myriad of interconnected systems in public service provision, have also responded reactively, leading to extended down-times and lengthy recovery periods that damage service provision [3]. However, although recovery from cyberattacks can lead to changes in IT policy and better-defended systems [5], these tend not to be accompanied by well-defined security-focused changes in user behaviors, a persistent issue evident through the last decade [8, 23, 24]: behavioral change tends to be led by IT departments and channelled out to personnel through policy communication and email directives, as an instruction-led set of best practices, and not as a set of principles underpinned by an objective to re-normalise user behaviors beyond the immediacy of the attack [6]. Failure to adequately address behaviors over the medium- to long-term can facilitate further attacks of a similar nature [8]. Allowing changed behaviors to wane over time can do the same. Whereas the technical aspects

of systems security can be managed methodically in many cases, it is the socio-behavioral that can enable persistence of threat to future cyberattack [10]. It follows that adequate address of user behaviors can, at a minimum, buttress organizational readiness, strengthen pre-emptive security policy, or help to identify a different set of normative behaviors by individuals commensurate with higher security. But beyond these impacts, a focus on user behaviors could provide a rebalancing of approach to cybersecurity to include human action as a core component to effective cyber *fitness*, and help organizations to move away from relatively chaotic reactive response to data breaches and system attacks, and towards a calmer and more planned proactive approach.

In many organizations, and indeed across supply chains, the suite of software applications, IT hardware and networking infrastructures in use span internal, dated, new and interlinked systems, some outsourced, some cloud-based, and at times with many different technical security measures in play [16, 17]. However, many breaches occur through on-premise local systems, typically where users have increased system privileges, or where systems are particularly dated with poor security controls [6]. Cloud-based systems, in the main, are relatively more secure when based on leading technologies such as AWS or Azure. The more preventative proactive approach to cybersecurity encompasses the technical, with ongoing focus on anti-malware and other software- and hardware-based security layers for IT systems, networks and underlying infrastructures. A focus on technically securing local on-premise applications, where persons interact with these systems daily, for example, health record entry in hospitals, or online database access, only goes part of the distance in better securing these and connected systems: user behaviors cannot be an afterthought [8]. Without clear attendance to behavior, cybersecurity is destined to repeat its own history, just perhaps with bigger stories.

III. User-Behavior- Phantom or Menace

It is important to scope what the information systems community of researchers already knows about user behaviors in the context of cybersecurity and, in particular, the impact *of* behavior *on* cybersecurity. This study aims to do exactly that, through a systematic literature review. However, even since the turn of the 21st century IS research has been desperately fragmented on appropriate methods, approaches, and implementations of research [18], with divergence persisting over more recent years with various approaches to systematic reviews [19]. Further, as cybersecurity has evolved in prominence and composition, literature reviews within the last half decade have already successfully provided increased clarity on various aspects of cybersecurity, for example blockchain [20], smart cities [21], and supply chain [16]. Additionally, a 2019 systematic literature review on the impact of user behaviors on cybersecurity focused specifically on time pressure [9], and spanning only 21 studies; whereas a 2021 review mentioned earlier focused particularly on a theoretical epistemic view, excluded papers from major IS conferences, did not study the broad range of user behavior, user psychology, social or corporate behavioral norms, and did not seek to identify behavioral themes [22]. The study here aims to a) employ a broader framework to expand upon these studies, to undertake a systematic literature review to map the research to date, focusing on the social rather than the technical aspects of cybersecurity within the information systems (IS) and informing literature bases, and b) to inform a second stage of research by identifying a set of common themes pertaining to user behaviors that warrant further study and address in proactive approaches to cybersecurity. Employing the procedural systematic ontology developed by [13] and following the model operationalised by [11], the study comprises a 3-stage structure. The first stage involves a review of research relevant to user behavior in cybersecurity, using the following 6-step linear funnelling process:

1. *Develop a review plan*
2. *Define and apply a literature search strategy*
3. *Select and refine studies*
4. *Assess and refine quality*
5. *Extract categories*
6. *Synthesise and classify findings*

The systematic approach taken provides "a simple, theoretically grounded, generalizable, yet flexible framework" [13], with a rigorous provision of both systematicity and transparency *of* and *within* method. For an exposition of both aspects, see [13]. However, though aspects of this approach have been employed in a small set of studies since 2016, a complete operationalisation in 2021 [11] provides a clear blueprint for execution in new domains. It is a cascading model, with steps 1-2 of 6 extensively dependent upon initial definition of search strategy, search terms, and bounds; rigor in these initial steps cascades through the model, with a further escalation in rigor required for step 5, which defines categories of concern in the particular domain. The 6-step approach may be buttressed by compositional modelling approaches such as prisma, to better categorise studies and identify categorical theme. We focus specifically on particular objectives to this paper, and to the first part of the tripartite study, in the following section.

In line with [11], the resultant theorising review will contribute to information systems research at the intersection of user behavior and cybersecurity. The output of the review is an exposition of shortcomings in the relevant extant literature, directly in line with [11], and explicitly cognisant of a-priori boundaries discussed in [12], in addition to the ranked identification of categories of primacy with respect to user behavior in cybersecurity. The theorising review will also provision the elaboration of extant knowledge in cyber-specific user behaviors and identify potential perils for organizations and users.

The second stage will forward a classification schema to investigate categories of primacy, identifying appropriate methodologies and approaches once categories are identified, and delineation of a set of further studies according to category. The third stage is the execution of these studies.

IV. Objectives

This paper focuses on the first stage, comprising a 6-step ontology for the literature review. Helping to identify sources of critical knowledge gaps, and missing and neglected aspects in research, a critical factor in effecting a

quality approach to systematizing a literature review using the 6-step funnelling process outlined above are, ab initio, to robustly assess and refine:

a) the sources of literature

b) the search terms employed, and

c) a rubric for categorization extraction.

In line with previous relevant systematic searches in information systems, starting proposals for literature sources are Web of Science and EBSCO (Business Source Premier, Scopus and PsychInfo), ACM Digital Library and IEEE Xplore. Additionally, to address the scoping shortcomings of previous systematic reviews , we propose to include AISEL Electronic Library to include premier IS conference proceedings such as ICIS, ECIS and HICSS; and the AIS senior scholar basket of 8 IS journals.

The systematic review will provide both theoretical and practical contributions for research and industry by (i) surfacing the epistemological conflicts in cybersecurity and user behaviors in extant literature, and (ii) creating a practical categorisation of cybersecurity user-behavior instances to various user- and organizational-specific contexts.

First, the review will provide one of the first examples of rigorous application of the dual systematicity and transparency approach to user behavior in cybersecurity, and in so doing, will both challenge and advance existing thinking in the domain.

The paper will apply a pluralist approach to categorize cybersecurity user behavior systematicity and their corresponding variances in scoping mechanisms. Pluralism, an approach encompassing an expansive epistemological research perspective typically involving multiple research methods and determination of what is 'real', underscores the crux of cybersecurity user behavior variance - stemming from how individual users' efficacies, values, beliefs, and worldviews shape their interpretation of mandated cybersecurity compliance policies. Consider a company with a password policy requiring "at least 14 characters with no repeated numbers or letters, at least five symbols and three upper case letters, non-dictionary words, and must be changed every fortnight, without repetition." Some users may interpret such a monistic rule as overly complex to remember and write down the password on a stick-it note under the keyboard. In the process, despite such a monistic policy, user pluralism can increase, rather than reduce, cybersecurity vulnerabilities. Karlsson et al. [24] confirms this dual systematicity and monistic-pluralistic conflict between cybersecurity compliance requirements and competing user values and priorities.

Monistic-pluralistic conflicts can lead to high-levels of cybersecurity user-behavior variance, signalling uncertainty. Findings from the review will help crystallize and underscore elemental behavioral precepts as best practices - identifying prudent, pluralistic rather than monistic "normative expectations" in society and industry.

Second, underpinned by pluralism, the paper will systematically categorise, classify, and tie instances of cybersecurity user-behavior failures to various contexts, with practical implications for research and industry. Our categorisation will be organized as an m x n matrix comprising of cybersecurity user behavior failures (m) and organizational/system context/processes (n). In our categorisation schema, *m* classifies the spectrum of cybersecurity user behavior failures - exposing organizational vulnerabilities. As shown in table 1, cybersecurity user behavior failures include malware downloads from visiting unknown or phished sites, failing to encrypt confidential data, leaked passwords from inadequate password management, privacy compromises by failing to secure personal information, among others.

Cybersecurity user behavior failures do not manifest themselves *in vacuo,* but occur in context of a particular organizational/user objective or organizational process, depicted as *n*. We use extant review literature to populate the m x n matrix to highlight cybersecurity user behavior failures for specific contexts, from access to culture. Finally, the populated matrix can allow us to infer and deduce a framework based on pluralistic challenges and well as "objective" elemental behavioral precepts as behavioral best practices.

TABLE I. BEHAVIORAL FAILURE MATRIX

Context/ Process (n)	**Cybersecurity User Behavior Failures (m)**			
	Malware Download	***Confidentiality (Encryption)/Privacy***	***Password Leakage***	**…**
Mobility/ Access	[3, 24]	[2, 5]	[5, 25]	
Productivity	[6, 24]	[4, 9, 11, 20]	[2, 9, 10]	
Efficiency/ Convenience	[2]	[4, 9, 17, 21, 25]	[24]	
…				

For example, users might download malware because users need to (i) use public networks for access (ii) download a program from an unsecure site to read a particular file to stay productivity [6, 24], (iii) download files/programs while travelling [3, 24]. Similarly, users might fail to follow strict encryption guidelines because users may need to (i) travel and use hotel, client or vendor computers to read files [2, 5], or (ii) share data across vendors and collaborators for efficiency/convenience [9, 25].

Subsequently, the study will contribute to normative expectations on user behaviors in cybersecurity, raise awareness of categories of behavior, and help organizations and managers to better understand the perils of particular behaviors by establishing more mindful cybersecurity standard operating procedures.

V. CONCLUSION

This paper describes the first stage of a 3-stage study on user behaviors in cybersecurity. The paper is focused on development of a systematic literature review following a particular rigorous approach based on the duality of monistic policies and pluralistic user behaviors. The study will be operationalised using the cybersecurity user behavior failures and context matrix to frame categorising themes. The study will surface whether cybersecurity user behavior failures are a phantom phenomenon or a menace that needs mitigation. The paper outlined the value and contributions possible through this review; and outlined the importance of initial rigour in scoping the review sources, defining search terms, and categorising themes. The initial study should also identify gaps in extant research, in particular epistemological imbalances, lack of attention to

particular user behavioral consequence, or scant consideration of cybersecurity issues of growing importance as the area evolves further. In so doing, the study hopes to help in illuminating human factors of growing import for cybersecurity reliance and efficacy.

REFERENCES

[1] L. Kappelman, V. Johnson, R. Torres, C. Maurer, and E. McLean, "A study of information systems issues, practices, and leadership in Europe," *European Journal of Information Systems,* vol. 28, no. 1, pp. 26-42, 2019/01/02 2019, doi: 10.1080/0960085X.2018.1497929.

[2] A. Bahuguna, R. K. Bisht, and J. Pande, "Assessing cybersecurity maturity of organizations: An empirical investigation in the Indian context," *Information Security Journal: A Global Perspective,* vol. 28, no. 6, pp. 164-177, 2019/11/02 2019, doi: 10.1080/19393555.2019.1689318.

[3] D. F. Norris, L. Mateczun, A. Joshi, and T. Finin, "Managing cybersecurity at the grassroots: Evidence from the first nationwide survey of local government cybersecurity," *Journal of Urban Affairs,* vol. 43, no. 8, pp. 1173-1195, 2021/09/14 2021, doi: 10.1080/07352166.2020.1727295.

[4] J. Lin, L. Carter, and D. Liu, "Privacy concerns and digital government: exploring citizen willingness to adopt the COVIDSafe app," *European Journal of Information Systems,* vol. 30, no. 4, pp. 389-402, 2021/07/04 2021, doi: 10.1080/0960085X.2021.1920857.

[5] C. Donalds and C. Barclay, "Beyond technical measures: a value-focused thinking appraisal of strategic drivers in improving information security policy compliance," *European Journal of Information Systems,* pp. 1-16, 2021, doi: 10.1080/0960085X.2021.1978344.

[6] W. A. Cram, J. G. Proudfoot, and J. D'Arcy, "Organizational information security policies: a review and research framework," *European Journal of Information Systems,* vol. 26, no. 6, pp. 605-641, 2017/11/01 2017, doi: 10.1057/s41303-017-0059-9.

[7] K. C. Ng, X. Zhang, J. Y. L. Thong, and K. Y. Tam, "Protecting Against Threats to Information Security: An Attitudinal Ambivalence Perspective," *Journal of Management Information Systems,* vol. 38, no. 3, pp. 732-764, 2021/07/03 2021, doi: 10.1080/07421222.2021.1962601.

[8] M. L. Jensen, A. Durcikova, and R. T. Wright, "Using susceptibility claims to motivate behavior change in IT security," *European Journal of Information Systems,* vol. 30, no. 1, pp. 27-45, 2021/01/02 2021, doi: 10.1080/0960085X.2020.1793696.

[9] N. H. Chowdhury, M. T. P. Adam, and G. Skinner, "The impact of time pressure on cybersecurity behavior: a systematic literature review," *Behavior & Information Technology,* vol. 38, no. 12, pp. 1290-1308, 2019/12/02 2019, doi: 10.1080/0144929X.2019.1583769.

[10] X. A. Zhang and J. Borden, "How to communicate cyber-risk? An examination of behavioral recommendations in cybersecurity crises," *Journal of Risk Research,* vol. 23, no. 10, pp. 1336-1352, 2020/10/02 2020, doi: 10.1080/13669877.2019.1646315.

[11] L. M. Giermindl, F. Strich, O. Christ, U. Leicht-Deobald, and A. Redzepi, "The dark sides of people analytics: reviewing the perils for organisations and employees," *European Journal of Information Systems,* pp. 1-26, 2021, doi: 10.1080/0960085X.2021.1927213.

[12] F. Rowe, "What literature review is not: diversity, boundaries and recommendations," *European Journal of Information Systems,* vol. 23, no. 3, pp. 241-255, 2014/05/01 2014, doi: 10.1057/ejis.2014.7.

[13] G. Paré, M. Tate, D. Johnstone, and S. Kitsiou, "Contextualizing the twin concepts of systematicity and transparency in information systems literature reviews," *European Journal of Information Systems,* vol. 25, no. 6, pp. 493-508, 2016, doi: 10.1057/ s41303-016-0020-3.

[14] A. Jeyaraj and A. H. Zadeh, "Exploration and Exploitation in Organizational Cybersecurity," *Journal of Computer Information Systems,* pp. 1-14, 2021, doi: 10.1080/08874417.2021.1902424.

[15] S. Yusif and A. Hafeez-Baig, "A Conceptual Model for Cybersecurity Governance," *Journal of Applied Security Research,* vol. 16, no. 4, pp. 490-513, 2021/10/02 2021, doi: 10.1080/19361610.2021.1918995.

[16] S. A. Melnyk, T. Schoenherr, C. Speier-Pero, C. Peters, J. F. Chang, and D. Friday, "New challenges in supply chain management: cybersecurity across the supply chain," *International Journal of Production Research,* pp. 1-22, 2021, doi: 10.1080/00207543.2021.1984606.

[17] A. Mahmud, "Application and criminalization of artificial intelligence in the digital society: security threats and the regulatory challenges," *Journal of Applied Security Research,* pp. 1-15, 2021, doi: 10.1080/19361610.2021.1947113.

[18] H. K. Klein, "Crisis in the IS field? A critical reflection on the state of the discipline," *Journal of the Association for Information Systems,* vol. 4, no. 10, 2003.

[19] M. Tate, E. Furtmueller, J. Evermann, and W. Bandara, "Introduction to the special issue: the literature review in information systems," *Communications of the Association for Information Systems,* vol. 37, no. 5, 2015.

[20] M. Liu, W. Yeoh, F. Jiang, and K.-K. R. Choo, "Blockchain for Cybersecurity: Systematic Literature Review and Classification," *Journal of Computer Information Systems,* pp. 1-17, 2021, doi: 10.1080/08874417.2021.1995914.

[21] G. Verhulsdonck, J. L. Weible, S. Helser, and N. Hajduk, "Smart Cities, Playable Cities, and Cybersecurity: A Systematic Review," *International Journal of Human–Computer Interaction,* pp. 1-13, 2021, doi: 10.1080/10447318.2021.2012381.

[22] T. Simon, "Revolution and stability in the study of the human factor in the security of information systems field : A systematic literature review over 30 years of publication," in *2021 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, 14-18 June 2021 2021, pp. 1-8, doi: 10.1109/CyberSA52016.2021.9478219.

[23] 95% of Successful Security Attacks are the Result of Human Error, *Security Magazine*, June 19, 2014. url: https://www.securitymagazine.com/articles/85601-of-successful-security-attacks-are-the-result-of-human-error. Last accessed: January 11th, 2022.

[24] F. Karlsson, M. Karlsson, and J. Åström, "Measuring employees' compliance – the importance of value pluralism," *Information and Computer Security*, Vol. 25 No. 3, pp. 279-299, 2017.

[25] P. Datta, "Hannibal at the gates : Cyberwarfare & the Solarwinds sunburst hack," J*ournal of Information Technology Teaching Cases*. March, 2021. 204388692199312. 10.1177/2043886921993126.

VICSORT - A Virtualised ICS Open-source Research Testbed

1st Conrad Ekisa, 2nd Diarmuid Ó Briain and 3rd Yvonne Kavanagh
engCORE Research Centre,
Department of Aerospace, Mechanical & Electronic Engineering,
Institute of Technology, Carlow, Ireland
Email: conrad.ekisa@itcarlow.ie

***Abstract*—Industrial Control Systems (ICS) are at the forefront of most, if not all the critical infrastructure and critical service delivery. ICS underpin modern manufacturing and utility processes and greatly contribute to our day-to-day livelihoods. However, there has been a significant increase in the number and complexity of cyberthreats specifically targetted at ICS, facilitated by increased connectivity in an effort to improve production efficiency. Furthermore, the barriers of entry to ICS cybersecurity are still high given the limited skills base, expensive and proprietary hardware and software as well as the inherent dangers of manipulating real physical processes. This greatly inhibits the practical application of cybersecurity tools in ICS environments and therefore the opportunity for practitioners to gain valuable ICS cybersecurity experience. ICS Testbeds are often either expensive and are not necessarily holistic enough to provide learners with the complete breadth of ICS. This paper introduces VICSORT, a open-source virtualised ICS testbed that provides a platform for ICS cybersecurity learners and practitioners to interface with an ICS environment that closely emulates a real-world ICS, as well as explore and practice techniques for attack and consequently defence of an ICS. VICSORT builds upon the Graphical Realism Framework for Industrial Control Systems (GRFICS) to offer an easier to deploy environment with greater flexibility, whilst requiring significantly less resources all reducing the cost to the learner.**

***Index Terms*—ICS, Cybersecurity, Open-source, Virtualised, Testbed**

I. Introduction

ICS, in some form, have existed since the 20th century and certainly long before the advent of the Internet and connectivity were as prevalent as they are today. At the time, the most common ICS security design principle was *security by obscurity*. This means that if a system is air-gaped, that is, contained with no outside connections, then the system is inherently secure. However, over the years, in an effort to provide more visibility, efficiency and information about the industrial processes, there is more connectivity both to, and within, the ICS zone as well as greater synergy between Enterprise and Control zones. Advantageous as this can be, new ICS cyber attack vectors are exposed. This, coupled with the fact that ICS were initially designed with a *security by obscurity* principle, has left many ICS vulnerable to attack by malicious actors. Due to the critical nature of a number of ICS, there is now a understanding of the need to drive towards building and growing ICS cybersecurity aptitude and competence to better protect these systems.

Funded by the 2021 Government of Ireland - International Education Scholarship

However, the barriers of entry to ICS cybersecurity are still high given the limited skills base, expensive and often proprietary hardware and software, as well as the inherent dangers of manipulating real physical processes. These factors serve to greatly inhibit new entrants to the ICS cybersecurity space and became the motivation underpinning this research work. Conti et al [1] describe three types of testbeds: physical, virtual and hybrid testbeds. The focus of this paper shall be on the VICSORT virtual testbed.

This paper wishes to acknowledge and recognise the work of the Georgia Institute of Technology and Fortiphyd Logic [2] in the development of the Graphical Realism Framework for Industrial Control Systems (GRFICS) testbed which provided a foundation upon which VICSORT is built. The GRFICS testbed made the following contributions to the ICS cybersecurity research space:

- Conversion of the simplified Tennessee Eastman (TE) Challenge process simulation into a more portable and accessible format.
- Novel 3D visualisation of a dynamic chemical process simulation to increase engagement and realism.
- The most complex and complete virtualisation of an ICS network to date, released Free, Libre and Open-Source Software (F/LOSS)
- Modular framework for easy expansion or conversion to other physical processes and protocols.

This paper presents VICSORT, a virtualised ICS open-source research testbed that builds upon the GRFICS testbed with the following contributions:

- An all-in-one F/LOSS based testbed rebuilt with GNU/LinuX Containers (LXC) to provide a leaner overall build requiring significantly less system resources to operate. The testbed is easily deployable offline or online on the Cloud.
- A detailed guide on how to reproduce the testbed from scratch to provide a more in-depth understanding to how the ICS components operate.
- A number of modifications to the testbed components including Python library upgrades/replacements to the currently outdated and publicly unavailable libraries used

in the GRFICS version, Operating System (OS) upgrades, firewall upgrades to more closely mimic a real-world ICS, as well as architectural and design modifications.

In previous work [3], a four-part methodology developed as an aid to visualise ICS cybersecurity weaknesses and test remediation strategies is documented; (i) Identification of a suitable ICS Testbed, (ii) An ICS Cyber Attack (iii) Development of an ICS evaluation and risk mitigation strategy and (iv) ICS cybersecurity Toolkit. This paper introduces VICSORT as a testbed to support this methodology. This will include the demonstration of various ICS focused cyber attacks and weaknesses as highlighted in (ii) as well as poist mitigation strategies in (iii).

The remainder of this paper is organised as follows: Section II provides a background on ICS giving an overview of the general architecture, ICS components as well as the common protocols used to link systems. Section III discusses related ICS testbed work as well as challenges with the GRFICS testbed that the VICSORT testbed addresses. Section IV discusses the VICSORT testbed describing its network topology, components involved as well as enhancements included in this testbed. Section V provides an evaluation of the testbed. Section VI highlights the contributions made by VICSORT as well as referencing where and how the testbed can be accessed. Section VII finally provides a conclusion to this paper.

II. RELATED WORK

A. Previous Testbeds

In previous work [3], six ICS Testbed models were highlighted from Thiago et al [4], Genge et al [5], Maynard et al [6], Giani et al [7], David et al [2] and Bertrand et al [8].

Furthermore, other testbeds such as Hui et al [9], Candel et al [10], Korkmaz et al [11] and Gardiner et al [12] have also been reviewed.

All these testbeds exhibit great individual strengths such as closely modelling ICS operations, integrating major ICS components, as well as virtualising ICS components for easier deployment and modification. However, similar constraints surround a number of them such as non-replicability due to non-publicly available source code, lack of a realistic visual element to the ICS physical process, requirement to purchase specific physical components to implement the ICS physical process, or lack of a holistic representation of an ICS from the testbed models.

Furthermore, there is notably insufficient publicly available documentation with regards to setup, deployment, testbed specifications and limitations, within other reviewed ICS testbeds such as [13], [14], [15] to enable learner reproducibility as well as testbed scaling and possibly application diversification. This paper attempts to bridge that gap and introduces VICSORT in its own right, noting its specification, design parameters and implementation limitations.

B. Testbed Technical Requirements

In an effort to produce a testbed model that is reproducible, easily deployable, scalable, lean in terms of compute resources, as well as offering a contribution towards practical hands-on ICS cybersecurity training, the following testbed technical requirements were set out [3]:

- The testbed integrates major OT components such as Programmable Logic Controllers (PLC), Human Machine Interfaces (HMI), engineering workstations, firewalls and a physical process.
- The testbed incorporates the Modbus/TCP ICS protocol given their prevalence in real-world ICS implementations today.
- The testbed architectural build closely follows the well known ICS Purdue model.
- The testbed is built using F/LOSS tools to facilitate ease of implementation, and offers potential for future enhancements.
- The testbed incorporates a 3D visualisation to simulate the ICS control process. This is to allow for visualisation of the physical consequences of a successful cyber attack on an ICS.
- The entire testbed build is compute resource lean. The resource constraints set for this project were 16GB RAM, 100GB HDD, 4CPUs @ 2.4GHz, 4GB dedicated graphics card.

C. GRFICS

As cited in previous work [3], David et al [2] propose the Graphical Realism Framework for Industrial Control Systems (GRFICS). Developed by researchers from Georgia Institute of Technology and Fortiphyd Logic, GRFICS [16] is a open source ICS simulation tool based on the TE process [17] with a goal of bringing practical ICS security skills to a wider audience. The testbed, built using Python on GNU/Linux, is currently designed for educational purposes and offers only a single ICS process, that is to say, the TE process. Fortiphyd Logic have built similar simulations, available commercially, on their training portal. The GRFICS platform is comprised of a total of five Virtual Machines (VMs), built on the Oracle VirtualBox hypervisor that perform the functions of; a PLC, a HMI, a firewall, an engineering workstation and a novel 3D visualisation of the physical process [16]. The testbed also includes a network setup to be implemented within the Oracle VirtualBox hypervisor. The PLC is implemented using *OpenPLC* [4], the HMI is implemented using *ScabaBR* [18], the firewall is implemented with *pfSense* [19], an engineering workstation that is a standard workstation with a GNU/Linux Operating System (OS) and the software to make changes to the PLC, and lastly, the 3D physical process simulation is implemented with Unity Game Engine [20]. The testbed supports the running of a number of ICS related attacks such as Man in the Middle (MitM) attacks, Command Injections, False Data Injection, PLC Reprogramming, Loading Malicious Binary Payloads, and common IT attacks such as password cracking. While the GRFICS testbed is prebuilt, the source-code is available and can be customised or modified to meet the user's requirements.

D. VICSORT contributions to GRFICS

VICSORT improves the fundamental GRFICS build to provide for a smaller, leaner and a more easily deployed testbed build. Specifically, VICSORT furthers the work achieved in GRFICS by providing:

- An all-in-one open-source testbed rebuilt with LXD container manager and Kernel-based Virtual Machine (KVM) hypervisor to provide a leaner overall build in terms of required compute resources, that's easily deployable locally or on a Cloud deployment.
- A publicly available how-to guide detailing how to reproduce the testbed to provide a more in-depth understanding to the testbed build process as well as how ICS components operate and work together.
- A number of modifications to the testbed components including Python library upgrades/replacements to the currently outdated and publicly unavailable libraries used in the previous version, OS upgrades, firewall upgrades to more closely mimic a real-world ICS, as well as architectural and design modifications.

III. VICSORT

VICSORT is a modified build of the GRFICS testbed that seeks to lower the barrier of entry to ICS cybersecurity even further as well as contribute to practical hands-on ICS cybersecurity training. VICSORT attempts to provide a holistic overview of an ICS environment integrating the major components of an ICS namely; the PLC, HMI, engineering workstation, firewall and physical process. VICSORT virtualises all these components and builds a network topology to facilitate any required communication between the participating nodes. The previous section introduced the contributions made by VICSORT and this section shall expound on these contributions in more detail. VICSORT, built on GNU/Linux OS, heavily leverages LXC with an LXD container manager, Python programming, and KVM for its implementation. While a fair understanding of these elements necessary to fully appreciate the VICSORT build process, this is not absolutely required to run the testbed. This testbed is designed to support ICS cybersecurity students, new entrants to the ICS cybersecurity space and industry practitioners looking to reskill or upskill in ICS cybersecurity.

This section shall discuss various aspects of the testbed including virtualisation in ICS, testbed architecture, testbed components and initialisation procedures.

A. Virtualisation in ICS

Due to its many advantages, such as lower hardware costs, more efficient energy consumption and increase resource efficiency, virtualisation became one of the most influential technologies of the last 10 years within the enterprise industry. In more recent years, virtualisation itself has been challenged by light-weight containerisation, though it is considered a complement to virtualisation rather than a replacement. Despite this sea change in enterprise computing, the adaption of virtualised or containerised infrastructure within ICS has been slow with the highest adaption being the virtualisation of the HMI.

Taigo et al [21] discuss ICT-like virtualisation technologies within ICS such as Network Function Virtualisation (NFV) and Software Defined Networking (SDN) and note a number of advantages as well as associated challenges within ICS such as latency overhead that may not be acceptable for real-time operation requirements within ICS. They add, that despite the constraints, the potential efficiency, security and reliability benefits for ICS are enough to justify the progressive development and introduction of domain-aware virtualisation technologies within ICS. VISCORT leverages virtualisation technologies, such as SDN and NFV among others, to achieve an all-in-one virtualised testbed. This paper does not study the per-component performance differences between virtualised components and their physical counterparts but applies ICS virtualisation techniques and notes that the advantages of ICT-like virtualisation could become a significant enhancement to the future of ICS.

B. Testbed Architecture

The testbed attempts to closely mimic the Perdue model whilst highlighting the major components of an ICS. The Purdue module divides a manufacturing company or utility network into six levels broadly categorised under the Enterprise, De-militarised and Control Zone. These levels were described in previous work [3].

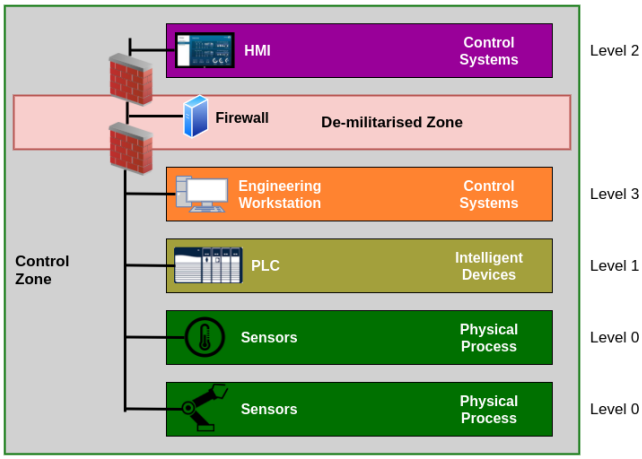


Fig. 1. VICSORT Purdue Model Implementation

This testbed focuses on the Control Zone, as depicted in Fig 1. This implementation highlights four levels in the Purdue model:

- *Level 0*: This includes sensors and actuators, usually termed as field devices, that interact with the actual physical processes, also found at this level.
- *Level 1*: This is the basic control level and is made up of PLCs, Remote Terminal Units (RTU) that control the physical devices.
- *Level 2*: This is the area supervisory control level and is typically comprised of HMIs at the shop or factory floor.

End-user or service engineers can interface with them. Alarm systems and control room workstations are also found at this level.

- *Level 3*: This is the site operations level that typically houses systems that supports plant control operations like a file sharing server, data historian, reporting and scheduling systems.

VICSORT v1.0 will focus on the Control Zone and may grow to incorporate the Enterprise zone in future iterations.

C. Testbed Network Topology

From this point forward, the baremetal computer or VM hosting VICSORT, will be termed as 'host-compute'.

VICSORT is divided into two networks namely the ICS or Control Zone network and the De-militarised Zone (DMZ) network. As illustrated in 2, the VICSORT network topology, the Control Zone is assigned to the 192.168.95.0/24 IP subnet and the DMZ is assigned to the 192.168.90.0/24 IP subnet. During normal testbed operation, communication between the two subnets and more specifically, communication between nodes is managed via the firewall. All traffic within the testbed is routed through the firewall. The testbed comprises of five nodes with a static IP address mapping as listed in Table I. The testbed also includes an attacker node that has successfully breached the DMZ Zone. This attacker node obtains an IP address via Dynamic Host Control Protocol (DHCP) on the DMZ network. However, the DHCP Server is not included in the setup but operates in the background.

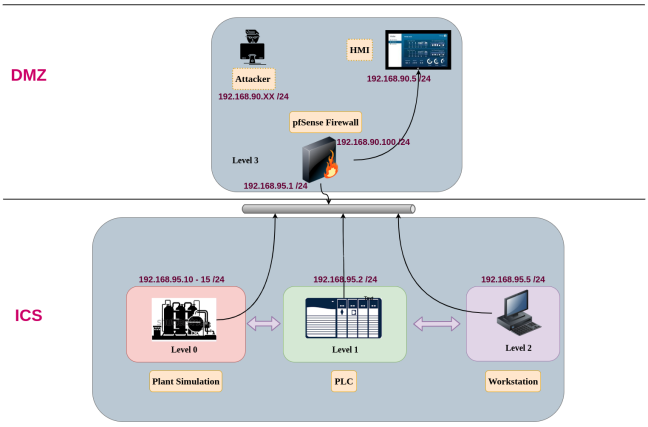


Fig. 2. VICSORT Network Topology

TABLE I
VICSORT IP ADDRESS MAPPING

Node	IP Address Mapping
HMI	192.168.90.5 /24
Firewall	- WAN: 192.168.90.100 /24 - LAN: 192.168.95.100/24
PLC	192.168.95.2 /24
Engineering Workstation	192.168.95.5 /24
Plant Simulation	192.168.95.10 - 15 /24
Attacker	192.168.90.XX /24

D. Testbed Components

As already outlined, the testbed comprises of a total of six components. This section will describe each of these components:

a) HMI: The HMI function is hosted in a LXC with an independent Ubuntu Linux 20.04 LTS OS. The HMI is implemented in software using *ScadaBR* [18] and resides within the DMZ – 192.168.90.0/24 network. During normal operation, the HMI is accessible from the host-compute via *http://192.168.90.5:9090/ScadaBR*. The HMI is referred to as **hmi-container** within VICSORT.

b) PLC: The PLC function is also hosted in a LXC with an independent Ubuntu Linux 20.04 LTS OS. The PLC is implemented in software using *OpenPLC* v2 [4] and resides within the Control Zone – 192.168.95.0/24 network. During normal operation, the PLC is accessible from the host-compute via *http://192.168.95.2:8080*. The PLC is referred to as **plc-container** within VICSORT.

c) Engineering Workstation: Additionally, the Engineering Workstation function is hosted in a LXC with an independent Ubuntu Linux 20.04 LTS OS. This node is specifically intended for the creation of PLC programs using ladder logic edited using *OpenPLCEditor* [22] and the completed programs can then be uploaded to the PLC. For this reason, this node is hosted on the same network as the PLC. This node resides within the Control Zone – 192.168.95.0/24 network. The Engineering Workstation node features a Graphical User Interface (GUI) to allow a user to interface with the node and create or modify PLC logic. The Engineering Workstation node is referred to as **workstation-container** within VICSORT.

d) Physical Process: The physical process is hosted in a LXC with an independent Ubuntu Linux 16.04 LTS OS. This node combines a physical process simulation that runs on the Unity Gaming Engine platform and virtualised sensors communicating using Modbus. Specifically, the simulation is implemented with Unity's WebGL that facilitates the simulation to be run via an Apache Tomcat web server hosted locally on the container.

The simulation is designed around the TE Challenge Process [17] and simulates a reactor core, input elements *A* and *B*, as well as a Purge output and the Product output. The TE Challenge Process is further discussed in the subsection III-F.

Consequently, six virtualised sensors that generate readings displayed within the simulation are also hosted on this node. These sensors, all implemented using the *pymodbus* Python library are *feed1.py*, *feed2.py*, *product.py*, *purge.py*, *analyzer.py* and *tank.py* with an IP address mapping as listed in Table II.

This node resides within the ICS Zone – 192.168.95.0/24 network. The simulation is accessible via *http://192.168.95.10*. This node is referred to as **simulation-container** within VICSORT.

e) Firewall: The firewall of choice for this testbed is *pfSense* [23] that runs on FreeBSD [24] OS. Within VICSORT, *pfSense* is hosted on a VM and is attached to the ICS and DMZ networks. The IP address mapping within the ICS zone is 192.168.90.100/24 and 192.168.95.100/24

TABLE II
PHYSICAL PROCESS IP ADDRESS MAPPING

IP Address	Port Bindings	Program Name	Function
192.168.95.10/24	502	Python	feed1.py
192.168.95.11/24	502	Python	feed2.py
192.168.95.12/24	502	Python	purge.py
192.168.95.13/24	502	Python	product.py
192.168.95.14/25	502	Python	tank.py
192.168.95.15/24	502	Python	analyzer.py
0.0.0.0	55555	Simulation	simulation server that updates 3D visualisation
0.0.0.0	80	Apache2	access the simulation via web browser

within the DMZ. The firewall has custom firewall rules present upon deployment via a *base_firewall_configs* file. The firewall implementation associates the ICS zone with the Local Area Network (LAN) and the DMZ with the Wide Area Network (WAN). The WAN is intended to face the Enterprise network whilst the LAN faces the Control zone.

The following rules on the WAN within the ruleset:

1) Allow all communication from the HMI to PLC
2) Allow access to the *OpenPLC* Web User Interface (UI) from the WAN network
3) Allow access from WAN to simulation VM web interface
4) Allow Internet access for all nodes on the WAN. This rule is disabled by default and should be enabled when Internet access is required on the WAN nodes for example for repository updates or further package downloads. However, the attacker node is exempt from this rule and always has Internet access.

The following rules, associated with the LAN are included within the ruleset:

1) Allow Internet Control Message Protocol (ICMP) to firewall from the 192.168.95.0/24 (LAN) network. This allows nodes on the LAN to ping the firewall
2) Allow all communication from the PLC to HMI
3) Allow Internet access to all nodes on the LAN. This rule is disabled by default and should be enabled when Internet access is required on the WAN nodes for example for repository updates or further package downloads.

All the nodes, except the attacker node, have their default route pointing to the firewall. Therefore, traffic from these nodes is all routed via the firewall. The *pfSense* web UI is accessible via *http://192.168.95.100*. This node is referred to as **pfsense-vm** within VICSORT.

f) Attacker: The attacker node integrated into VICSORT is Kali Linux 2021 running in an LXC. This node contains the entire Kali Linux suite of tools as well as a GUI accessible via Remote Desktop Protocol (RDP). This node resides in the DMZ and obtains an IP address dynamically from a DHCP server reading within that network. The assumption is that the attacker has successfully breached the enterprise network and managed to pivot to the DMZ zone. This node is referred to as **attacker-container** within VICSORT.

E. Testbed Underlying Build Design

The host-compute housing VICSORT is built based on the following criteria:

a) Internal Design: VICSORT relies heavily on LXC/LXD and KVM for its implementation. There are a total of five LXC and one VM in this implementation. The foundational technologies used in the development of VICSORT are Python and GNU/Linux. The networking between all these nodes is handled by the LXD networking function that creates two network bridges within the host-compute in accordance to the VICSORT network topology namely *icszone* and *dmzzone*. To maintain a low OS resource utilisation, VICSORT employs *lubuntu-desktop* [25] that is a fast and light-weight GUI for Ubuntu. The testbed also contains initial setup files available within the project Github repo that are necessary during the testbed's setup.

b) Intended Usage: This testbed is designed to be self-contained. This means that any Universal Resource Locaters (URL) used within the tested, for example the HMI URL, or URL to the simulation are intended to be accessed within the host machine. This testbed is designed to be suitable for deployment within a Type 2 hypervisor such as VirtualBox or VMWare, or on any cloud infrastructure. It should be noted however that these URLs and testbed resources can be accessed outside the host machine though this is not covered within the scope of this project. It is recommended to use one of these two options, either a Type 2 hypervisor or Cloud to host VICSORT as opposed to hosting the testbed directly on a Personal Computer (PC).

The host-compute contains a GUI to facilitate access to the various components of the testbed, many of which are accessible via a web browser. If the host-compute is within the Cloud, a user is required to RDP from their computer to the host-compute. Within this session, the user will be able to use the web browser to access web portal utilities that most components leverage for their operation. If the host-compute is within a Type 2 hypervisor, a user can simply utilise the console interface provided by the Type 2 hypervisor itself to access the host-compute, as opposed to using RDP. However, RDP is still an option in this case. Furthermore, the attacker container will also be accessed via RDP as its envisaged that the attacks will be initiated from the attacker node and not the host-compute. Figure 3 is a schematic illustrating how access to VICSORT, and the nodes within it, is managed.

c) Testbed Initialisation: The testbed contains a number of nodes. This warranted the need for an initalisation script to bring the testbed to an operational state upon boot of the host-compute. Upon boot of the host-compute, LXD is in a running state as a *systemd* service. The LXCs in most cases will all be in a running state too. The initialisation script initialises the testbed performing the following functions in chronological order.

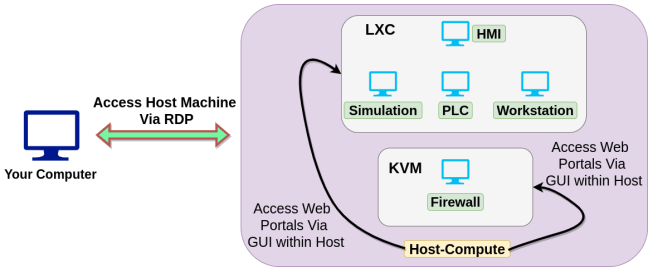


Fig. 3. Accessing the GUI components of VICSORT

The script obtains IP addresses for all the LXC. These IP addresses are static; however, they still need to be requested from the respective DHCP server. The script will then output all the containers and their corresponding IP addresses, start the firewall VM within KVM, provision the HMI and start *ScadaBR*, provision the PLC and start *OpenPLC*, provision the simulation node, start the simulation and pymodbus sensors, set the correct timezone within all the nodes, and finally configure a new default route on all the nodes except the attacker, pointing all their traffic to the firewall.

At this stage, a user is ready to start using the testbed.

F. Tennessee Eastman Challenge Process

The TE Challenge Process is based off the TE plant. Figure 4 illustrates the TE plant that the presents the challenge to establish control over.

The TE plant wide Industrial Control Process problem was proposed by Downs and Vogel in 1993 [26] as a challenge test problem to several control related topics such as multivariable controller design, optimisation, adaptive and predictive control and non-linear control. Since its design, over 60 studies have used this case study for alternative plant-wide control, process monitoring, fault detection and identification. The TE process is a realistic simulation environment of a real chemical process that is comprised of an exothermic two-phase (liquid and vapour) reactor, a flash separator, recycle compressor, and a reboiled stripper.

From Figure 4, the gas reactants A, C, D and E enter the reactor where the reactants undergo an irreversible exothermic (which means the temperature gradually increases) catalytic gas phase reaction. Since the reaction is exothermic in nature, it is cooled by the Cold Water Supply (CWS) and Cold Water Reset. Inert gas does not undergo chemical reactions under a set of given conditions. The partial condenser recovers the products from the reactor exit gas stream. The stripper is used to minimise the loss of reactants D and E in the liquid product stream. The gas overhead from the stripper is combined with the compressed overhead from the separator and recycled back to the reactor. The purge stream is used to prevent build-up of excess reactants, the inert B and the by-product F. The process produces two products from four reactants. Also present are an inert and a by-product making eight components A, B, C, D, E, F, G and H. The process variables are temperature, pressure, level and flow rate.

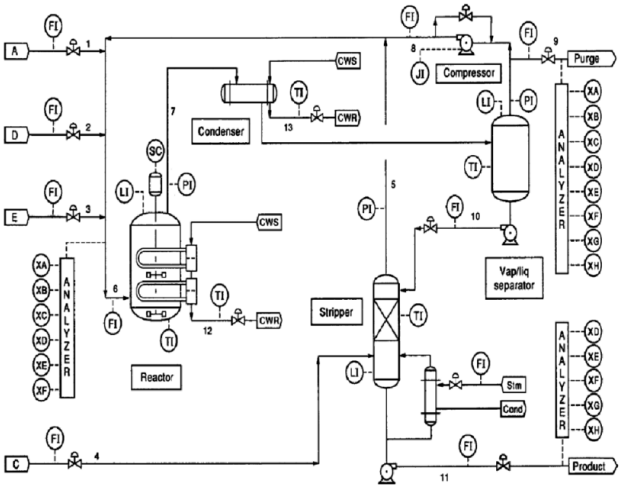


Fig. 4. Tennessee Eastman Challenge Process [26]

VISCORT simplifies these process with only two inputs; *feed1* and *feed2* as well as a Product output and a Purge byproduct. The process variables measured are pressure (kPa) and Level (%) and flowrate (kMol/h). Figure 5 illustrates the physical process simulation as well as its representation as shown on the HMI.

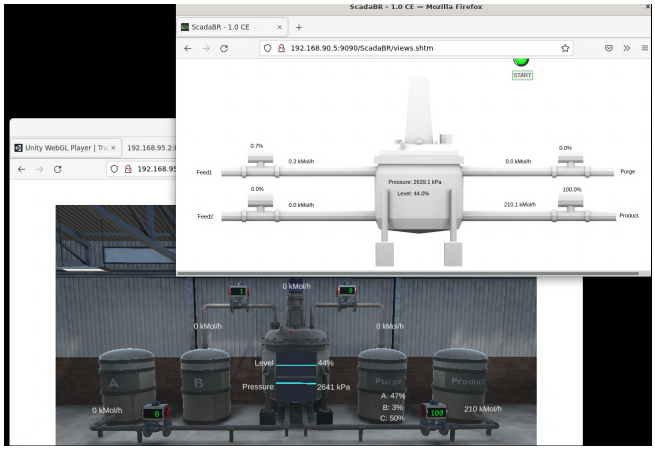


Fig. 5. HMI and Physical Process Simulation

G. Testbed in Operation

Figure 6 lists the five LXC and single KVM VM that make up VICSORT, in an operational state. Figure 5 illustrates the physical process simulation as well as its representation as shown on the HMI.

IV. BENCHMARKS

This section documents system benchmarks achieved by the testbed.

VISCORT implementation. This is a reduction from the 5GB minimum RAM requirement in GRFICSv2.
- VICSORT shrinks the required hard disk space required by the host-compute to about 23GB for an operational implementation. This is a reduction from the 38.69GB required in GRFICSv2.
- VICSORT makes OS updates utilising more recent OS implementations than GRFICSv2.
- VICSORT provides an F/LOSS light-weight build that is easily deployable on both a Type 2 hypervisor or a Cloud implementation.
- VICSORT provides a detailed installation guide to support learners that wish to replicate and build upon this deployment.

Current limitations to VISCORT, as inherited from GRFICS, include the lack of ease to the scaling of the number of sensors included within the testbed as this is closely tied to the 3D simulation. The testbed's data is generated by virtual sensors with the help of pymodbus [28]. The Modbus sensors in themselves can easily be scaled and the PLC modified to handle this change. However, the 3D simulation would require redesign and unfortuntately, the 3D simulation base files are not publicly available to facilitate this.

VICSORT is available at *https://gitlab.com/ekisac10/vicsort*. Included in this repository are VICSORT's required setup files, testbed scripts and setup manual. A VirtualBox *.ova* file is also available to support rapid implementation of the complete testbed on a VirtualBox VM.

VI. Conclusion and Future Work

This paper has introduced VICSORT – a light F/LOSS ICS testbed solution intended to be repeatable, scalable and easy to deploy. This is in an effort to lower the barriers to entry to ICS cybersecurity and provide a playground to further the ICS cybersecurity body of knowledge. This tool is licensed under the European Union Public Licence version 1.2 (EUPLv1.2) [29] and is currently in its first iteration (v1.0).

This paper wishes to acknowledge and recognise the work of the Georgia Institute of Technology and Fortiphyd Logic [2] in the development of the GRFICSv2 from which VICSORT builds upon.

For future works, it is intended that VICSORT will incorporate an Enterprise component as described by the Purdue model. The objective is to demonstrate the synergies between the Enterprise and Control zones as well as demonstrate how Information Technology (IT) cybersecurity affects Operational Technology (OT) and how IT and OT cybersecurity can complement each other.

References

[1] M. Conti, D. Donadel, and F. Turrin, "A Survey on Industrial Control System Testbeds and Datasets for Security Research," no. February, 2021. [Online]. Available: http://arxiv.org/abs/2102.05631

[2] D. Formby, M. Rad, and R. Beyah, "Lowering the Barriers to Industrial Control System Security with GRFICS," Georgia Institute of Technology and Fortiphyd Logic, Baltimore, MD, Tech. Rep., 2018. [Online]. Available: https://www.usenix.org/conference/ase18/presentation/formby

[3] C. Ekisa, D. Briain, and Y. Kavanagh, "An open-source testbed to visualise ics cybersecurity weaknesses and remediation strategies – a research agenda proposal," in *2021 32nd Irish Signals and Systems Conference (ISSC)*, 2021, pp. 1–6.

[4] T. Alves and T. Morris, "OpenPLC: An IEC 61,131–3 compliant open source industrial controller for cyber security research," *Computers and Security*, vol. 78, pp. 364–379, sep 2018.

[5] B. Genge, C. Siaterlis, I. Nai Fovino, and M. Masera, "A cyber-physical experimentation environment for the security analysis of networked industrial control systems," *Computers and Electrical Engineering*, vol. 38, no. 5, pp. 1146–1161, 2012.

[6] K. Mclaughlin and S. Sezer, "An Open Framework for Deploying Experimental SCADA Testbed Networks," *SCADA Cyber Security Research*, pp. 92–101, 2018. [Online]. Available: https://doi.org/10.14236/ewic/ICS2018.11

[7] A. Giani, G. Karsai, T. Roosta, A. Shah, B. Sinopoli, and J. Wiley, "A testbed for secure and robust SCADA systems," *ACM SIGBED Review*, vol. 5, no. 2, pp. 1–4, jul 2008.

[8] M. Bertrand and T. Olivier, "Dissertation - Simulating Industrial Control Systems Using Mininet," 2017. [Online]. Available: https://dial.uclouvain.be/memoire/ucl/en/object/thesis%3A14706/datastream/PDF_01/view

[9] P. Maynard and K. Mclaughlin, "Ics interaction testbed: A platform for cyber-physical security research," 2019.

[10] R. Candell, K. Stouffer, and D. Anand, "A cybersecurity testbed for industrial control systems," 10 2014.

[11] E. Korkmaz, A. Dolgikh, M. Davis, and V. Skormin, "Ics security testbed with delay attack case study," 11 2016, pp. 283–288.

[12] J. Gardiner, B. Craggs, B. Green, and A. Rashid, "Oops i did it again: Further adventures in the land of ics security testbeds," 11 2019, pp. 75–86.

[13] F. Sauer, M. Niedermaier, S. Kießling, and D. Merli, "Licster – a low-cost ics security testbed for education and research," 10 2019.

[14] J. Gardiner, B. Craggs, B. Green, and A. Rashid, *Oops I Did it Again: Further Adventures in the Land of ICS Security Testbeds*. ACM Press / Sheridan, Nov. 2019, p. 75–86.

[15] W. Xu, Y. Tao, C. Yang, and H. Chen, "Msicst: Multiple-scenario industrial control system testbed for security research," *Computers, Materials Continua*, vol. 58, pp. 691–705, 01 2019.

[16] D. Formby, "GitHub - djformby/GRFICS: Graphical Realism Framework for Industrial Control Simulations," 2018. [Online]. Available: https://github.com/djformby/GRFICS

[17] T. J. McAvoy, "BASE CONTROL FOR THE TENNESSEE EASTMAN PROBLEM," *Computers Chem Engng*, vol. 18, no. 5, pp. 383–413, 1994.

[18] M. Sistemas, "Release ScadaBR 1.2 · ScadaBR/ScadaBR · GitHub." [Online]. Available: https://github.com/ScadaBR/ScadaBR/releases/tag/v1.2

[19] P. Kamal, "Intrusion Detection using pfSense - Open Source FREEBSD Firewall," Tech. Rep., 2014. [Online]. Available: https://www.academia.edu/17560234/INTRUSION_DETECTION_USING_PFSENSE_OPEN_SOURCE_FREEBSD_FIREWALL

[20] U. Technologies, "Unity Real-Time Development Platform — 3D, 2D VR AR Engine." [Online]. Available: https://unity.com/

[21] T. Cruz, R. Queiroz, J. Proença, P. Simoes, and E. Monteiro, "Leveraging virtualization technologies to improve scada ics security," *Journal of Information Warfare 1445-3312 (Printed) ISSN 1445-3347*, vol. 15, 10 2016.

[22] T. Alves, "Openplc editor," 2018. [Online]. Available: https://www.openplcproject.com/plcopen-editor/

[23] E. S. Fencing, "pfsense," 2021. [Online]. Available: https://www.pfsense.org/

[24] "The freebsd project." [Online]. Available: https://www.freebsd.org/

[25] lubuntu Meilix, "lubuntu – lightweight, fast, easier." [Online]. Available: https://lubuntu.net/

[26] J. J. Downs and E. F. Vogel, "A plant-wide industrial process problem control," vol. 17, pp. 245–255, 1993.

[27] D. Formby, "GitHub - Fortiphyd/GRFICSv2: Version 2 of the Graphical Realism Framework for Industrial Control Simulation (GRFICS)," 2020. [Online]. Available: https://github.com/Fortiphyd/GRFICSv2

[28] H. Petrak, "Pymodbus - a python modbus stack — pymodbus 2.5.0 documentation." [Online]. Available: https://pymodbus.readthedocs.io/en/dev/readme.html

[29] G. Oettinger, "European union public lisence (eupl)," *Official Journal of the European Union*, pp. 1–6, 5 2007.

The Application of Reinforcement Learning to the FlipIt Security Game

Xue Yang
School of Computer Science
NUI Galway
Galway, Ireland
x.yang6@nuigalway.ie

Enda Howley
School of Computer Science
NUI Galway
Galway, Ireland
ehowley@nuigalway.ie

Michael Schukat
School of Computer Science
NUI Galway
Galway, Ireland
michael.schukat@nuigalway.ie

***Abstract*—Advanced Persistent Threat is currently one of the most important threats to industries and governments. It is used to describe an attack campaign in which intruders can persistently and stealthily compromise a sensitive resource. APT has proven to be difficult to detect and defend against in the cloud-based environment by traditional methods, calling for more advanced security technologies. FlipIt is a two-player security game where an attacker and defender compete to control a sensitive resource in advanced scenarios such as APTs. Its robustness against APT attacks is outstanding. We model the FlipIt game as a Markov Decision Process and apply reinforcement learning to the framework. The goal is to find an optimal adaptive strategy for a player to compete against any unknown opponent in a FlipIt game with incomplete information. This means the best result for a player is to maximize the ownership of the resource with minimum cost. We perform experiments on single-agent and multi-agent scenarios, respectively. We further extend the model to involve noisy information and consider the openness of the game. Our experimental analysis proves that in a two-player FlipIt game, an adaptive player can automatically learn and find an optimal strategy using only the last move information of the opponent, who moves with a non-adaptive strategy (i.e. a periodic strategy with random noise). The parameters related to the random noise we considered affect the average benefit for each player. In addition, we consider the openness of the game in which new participants are introduced individually at random time steps with a certain probability. In this case, the model is generalized from two-player to n-player, and the convergence of the optimal strategy learned by each player is confirmed. Moreover, we demonstrate that varying the probability of adding an additional player does not affect the convergence but changes the average benefits for players.**

***Index Terms*—Security Games, Advanced Persistent Threats, FlipIt, Reinforcement Learning, Adaptive Strategy, Random Noise, Game Openness**

I. Introduction

Industries and governments are becoming increasingly reliant on clouds because of the convenient and cost-effective access to distributed servers and shared resources [1]. However, due to the distributed nature of cloud computing, the vulnerability of cloud environments is an area of growing concern. The hacks and attacks to the cloud may lead to different types of damages to networks, devices and data, causing huge amounts of financial loss and leakage of data privacy. Recently, Advanced Persistent Threat (APT) is becoming one of the largest threats to companies and governments [2]. APT is used to describe an attack campaign in which intruders can persistently compromise a sensitive resource. It leverages "zero-day exploits" meaning that the vulnerability is publically unknown [3]. Moreover, it has a high degree of stealth that cannot be detected immediately by the defender [4]. The traditional signature-based security methods have proven to be ineffective in detecting and defending against APTs [3], introducing a need for more advanced security technologies.

The research of game theory in cloud security is promising as it provides a mathematical approach to modeling and analyzing complex cloud security problems [5]. A security game studies the interactions between attackers and defenders who attempt to maximize their objectives. So far, there have been several game-theoretic solutions proposed to address network security issues. Particularly, the FlipIt game which was introduced by [6] has strong robustness against APT attacks. It models the interactions between an attacker and defender who compete to control a sensitive resource completely under advanced scenarios such as APTs. FlipIt has a unique "stealthy" nature that a player does not know who is controlling the resource until he moves [7], which distinguishes FlipIt from other security games.

The basic FlipIt framework has the potential to be extended to variants that are applicable to different security scenarios [4]. The newly proposed QFlip is a Reinforcement Learning (RL)-based FlipIt game that models the interactions between an attacker and defender as a Markov Decision Process using the Q-learning algorithm (a famous Temporal-Difference RL algorithm) [7]. It learns the optimal adaptive strategy for an agent against an opponent playing with a range of non-adaptive strategies (e.g. periodic strategy, exponential strategy) and demonstrates effective results.

RL models the process that one or multiple agents interact with a dynamic environment and learn the optimal strategies automatically through trial-and-error. In this paper, motivated by the fact that RL is highly capable of modeling real-world stochastic games with imperfect and incomplete information, we dive deep into this area, exploring and evaluating RL methodologies in more complex FlipIt game including single-agent and multi-agent systems, and involving the uncertainty and openness of the game. An RL-based FlipIt game in a cloud data centre is demonstrated in Fig. 1.

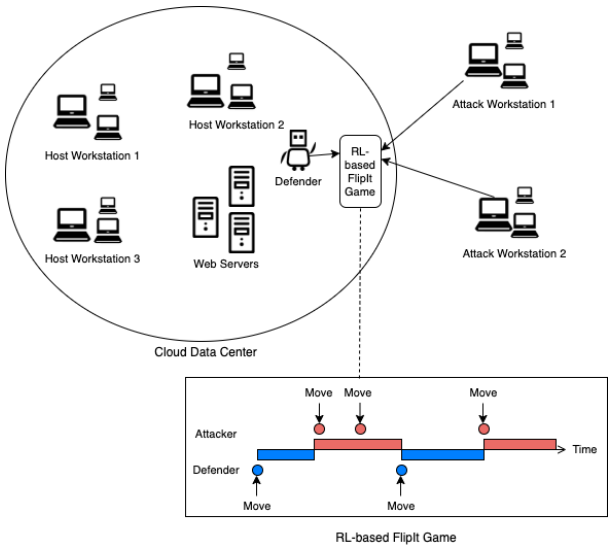


Fig. 1. Reinforcement Learning-based FlipIt Game in Cloud Data Center ([8], [9])

Our analysis of different FlipIt variants based on simulation experiments makes the following contributions. Firstly, we introduce random noise to a periodic player's strategy in a two-player RL-based FlipIt model, and verify that the adaptive agent is able to use only the last move (LM) information of the non-adaptive opponent to automatically learn and find an optimal strategy. The parameters related to the random noise (i.e. the probability of adding a noise, the scale of the noise) affect the average benefits of players. Secondly, we consider the openness of the game in which new participants are introduced at random time steps, respectively, with a certain probability. In this case, the model is generalized from two-player to n-player, and the convergence of the LM-adaptive strategy learned by each player is confirmed. The game thereby obtains a Nash equilibrium. Furthermore, varying the probability of adding new participants does not affect the convergence but changes the average benefits of players. Generally speaking, for the original defender and attacker (i.e. the fixed two players before new participants joined), the larger the probability, the lower the average benefit. We believe our work can reveal that RL is a promising technology for designing optimal adaptive strategies in complex and dynamic security games.

The remainder of this paper is organized as follows. We start with a comprehensive literature review to analyze related work in Section II. Then we introduce the background knowledge in Section III. We describe the methodologies and perform experimental results in Section IV and Section V, respectively, and finally conclude in Section VI.

II. Related Work

Game-theoretic methods have been extensively explored in the security area. Reference [6] introduced FlipIt as a novel two-player game where an attacker and defender compete to control a shared resource completely. They demonstrate the stealthy nature of FlipIt that the controller of the resource is unknown to a player until he moves. They restrict the defender with non-adaptive strategies and vary the attacker's strategy from periodic to adaptive. Subsequently, the applications of FlipIt including password reset policies, key rotation, VM refresh and cloud auditing were introduced by [4]. So far, many variants of FlipIt framework have been proposed: [10] extends the game to which a player can check the state (who is controlling the resource) before making a move; [11] generalizes FlipIt to multiple resources; and [12] proposes a model in which an attacker incrementally and stealthily takes ownership of the resource until finally comprising it completely. However, the above studies only focus on non-adaptive strategies which limits the capability of the game. Reference [13] was the first to consider adaptive strategies in FlipIt game but the defender is non-stealthy and the moves of the attacker are not-instantaneous. Recently, [7] and [14] apply RL technologies to the FlipIt game and focus on adaptive strategies.

Machine Learning (ML) is getting increasing attention in the field of cloud security. As stated in [15], various ML algorithms have been studied to overcome cloud security issues, e.g. SVM for secure cryptosystems, KNN for privacy-preserving, Naive Bayes for intrusion detection, etc. In recent years, RL (a ML paradigm) has merged in this area. Generally, an RL agent learns a policy through interactions with a dynamic environment with limited or even no prior knowledge. Considering that RL is highly capable of modeling real-world stochastic games with incomplete information, we are motivated to draw on the idea of using RL in advanced security games. Reference [16] designs an algorithm that can automatically find the vulnerable spots of an internet data center power system based on RL. Reference [17] studies how to use RL to automatically identify vulnerable spots in software-defined networking (SDN)-enabled cloud data center networks. In case of the FlipIt game, [7] introduces a QFlip strategy by combining Q-learning algorithm and FlipIt, which plays optimally against any opponent with incomplete prior knowledge. Reference [14] uses Deep Q-Networks (DQN) algorithm (a combination of deep learning and Q-learning) to solve the complex game and generalizes it to n-player. Our paper extends the existed research with the introduction of random noise to a non-adaptive player's strategy in an RL-based FlipIt model. Furthermore, we consider the openness of the game that new participants can be introduced during the game, therefore, generalizing it from two-player to n-player. It is important to note that the noisy information and openness are both controlled by particular parameters.

III. Background

The background knowledge related to this paper includes FlipIt Game, Markov Decision Process and Reinforcement Learning.

A. FlipIt Game

FlipIt is a two-player game where an attacker and defender compete to control a resource by performing moves with a specific strategy, respectively, which is also called "the game of Stealthy Takeover" [6]. The resource could be a password, an infrastructure, a secret key or other sensitive resource that is crucial to an organization. For both of the players in a game, they can move at any time at a cost. The objective for each player is to control the resource as long as possible with a minimized cost. The distinctive "stealthy" feature of the FlipIt game indicates that a player does not immediately know the current ownership of the resource or when the other player moves, and players can only perceive this until they move. In addition, a player's strategy determines how to move under a specific state and whether can "win" the game or not, and a strategy can be evaluated using benefits which are calculated based on the accumulated ownership of the resource and the total move cost during the game.

A strategy in FlipIt has two main categories: non-adaptive and adaptive. In case of a non-adaptive strategy, players do not receive any feedback during the game and the strategy can be determined before the game starts. Renewal strategy is a subclass of non-adaptive strategy which generates intervals between consecutive moves by a renewal process. The intervals are independently and identically distributed random variables. Examples of renewal strategies include periodic strategy, exponential strategy, uniform strategy, and normal strategy. This paper uses periodic strategy as an example of non-adaptive strategy where the interval between consecutive moves is a constant value. In case of an adaptive strategy, players receive feedback during the game and can adaptively adjust their moves. The feedback can be the last move time of the opponent (LM Strategy) or the full history of the opponent's move times (FH Strategy). This research employs LM Strategy as it is more complex and challenging. It is notable that the last move of the opponent known to the player maybe not be the actual last move due to the stealthy character of the FlipIt game.

B. Markov Decision Process

Markov Decision Process (MDP) is used for sequential decision making under uncertainty [18]. It is equipped with Markov Property that the next state s' only depends on the current state s. A MDP relies on a probabilistic model which can be represented by a tuple of $< S, A, T, R >$, consisting of a set of states, a set of actions, a transition function $T(s, a, s')$ that can model the environment, and a reward function $R(s, a, s')$. A policy π represents a solution to an MDP problem, and the goal is to find the optimal policy π^* which has the highest expected discounted sum of rewards. The value of a policy over a time horizon (infinite) can be represented as (1):

$$\sum_{t=1}^{\infty} \mathbf{E}[\gamma^t R(s_t, a_t, s_{t+1})] \quad (1)$$

where γ is a discounted factor ranging from 0 to 1.

In addition, the value of a state can be formulated using Bellman Equation which is the immediate reward of the current state plus the expected discounted value of the next state that is represented in (2).

$$V(s) = R(s) + \gamma max_{a \in A(s)} \sum_{S'} T(s, a, s')V(s') \quad (2)$$

C. Reinforcement Learning

RL is a tool to find the optimal policy for a MDP problem. It models the process in which an agent iteratively interacts with a dynamic environment and learns the optimal policy (if it exists) automatically through exploration and exploitation with little or no prior knowledge of the environment. At each iteration, the agent selects an action based on what he perceives from the environment. The environment updates accordingly and sends the next state and a reward (determined by a reward function) to the agent. The goal of the agent is to find an optimal policy that yields the highest accumulative reward in a long term. The balance between exploration and exploitation is important because it is related to whether the learned policy is really optimal or sub-optimal. In addition, Multi-Agent RL (MARL) describes the situation that multiple agents act in the same environment over time and interact to improve their own strategies within the architecture of RL. Each player's decision is influenced by other players' actions directly or indirectly via the state of the environment. The state of the environment transits based on the joint action of agents and the rewards are based on the joint action and states.

Q-learning algorithm is a model-free, temporal-difference RL algorithm that uses an action-state value function to learn the value of an action a in a particular state s, which is called Q(s,a) value. Q(s,a) is estimated from an immediate reward and the discounted estimated future reward of the best action in the next state, and is stored in a Q-table. A trade-off between exploitation and exploration should be considered following a specific strategy, e.g. $\epsilon - greedy$ exploration strategy. Section IV demonstrates more details of how Q-learning algorithm is combined in our model architecture.

IV. Methodology

We model FlipIt as a MDP and apply the LM-adaptive strategy based on the Q-learning algorithm. It is notable that in all of our experiments, we only consider the discrete and infinite version of FlipIt that $t \in \{0, 1, 2, ...\}$, and we assume a defender has some priorities over an attacker such as:

- At t=0, the defender has the ownership of the resource;
- An attacker pays a higher cost for a move than a defender;
- If an attacker and defender move together the defender will control the resource, indicating that the attacker pays the cost but does not gain additional time of ownership.

A. Modeling FlipIt as an MDP

In a basic FlipIt game, there are two players: a defender p_0 and an attacker p_1. Their costs for performing a move are k_0 and k_1, respectively, where $0 < k_0 < k_1$. We first begin by introducing p_0 to play periodically and p_1 to play with an

LM-adaptive strategy. At each time step $t \in \{1,\ 2,\ 3, \ldots\}$, the agent p_1 chooses an action a_t from the action set $A = \{0, 1\}$ (0: no-move, 1: move). Specifically, there are two kinds of move type: $flipping$ and $consecutive$. $Flipping$ indicates that a player grabs the resource from the opponent successfully by performing a move, while $consecutive$ means a player continues to move when it has already controlled the resource. Afterwards, the environment updates accordingly and sends the agent with a reward r_t and a new state s_{t+1}. The critical step is how to model the state and reward.

State Observation. Fig. 2 demonstrates an example of a FlipIt game consisting of an adaptive attacker and a non-adaptive defender. LM_0 and $LM_0{}'$ are the true LM of the opponent and the known LM perceived by the adaptive attacker, respectively. The state to the agent is defined as the time since the opponent's last known move at time step t: $s_t = t - LM_0{}'$. Due to the stealthy character of the game, $s_t \geq t - LM_0$.

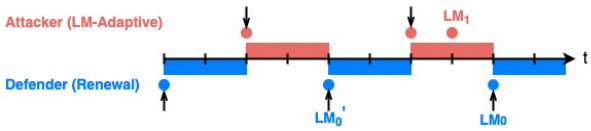


Fig. 2. the State to an Agent at Time t

According to [7], the state of the environment observed by the agent is depicted in Table I. It is important to note that in the case of $a_t = flipping$, LM_0 becomes the last known move of the opponent at the next time step, therefore, $s_{t+1} = t - LM_0 + 1$.

TABLE I
MOVE TYPES AND EXPLANATIONS

Action a_t	Cost	Outcome	Explanation
1 (flipping)	k_1	$s_{t+1} = t - LM_0 + 1$	$LM_0 > LM_1$ p_1 takes control
1 (consecutive)	k_1	$s_{t+1} = s_t + 1$	$LM_0 \leq LM_1$ p1 moves while in control
0 (no-move)	0	$s_{t+1} = s_t + 1$	p_1 does not move

Reward Function. It is simple to define a reward to the action type of $no - move$ or $consecutive$, which should be $-k_1$ and 0, respectively. However, it is challenging to find a reward when the action is $flipping$. [7] has proposed a reward function to the action type of $flipping$ that has been experimentally proven highly effective:

$$r_t = \frac{\rho - k}{c} \tag{3}$$

where ρ is the estimated average move frequency of the opponent, k is the agent's cost per move, and c is a normalization constant. It indicates that when the action type is *flipping*, the reward r_t from the environment is a fixed constant that is related to the agent's move cost and the opponent's move frequency. The value of ρ can be estimated by starting the game and keeping track of the observed move time ticks of the opponent. However, this reward function is only valid in a single-agent system rather than a multi-agent system. Because in a multi-agent game in which all the players adopt a LM-adaptive strategy, they do not have a fixed constant to indicate the average move frequency. To solve this problem, we define a more general reward function in which at time step t, the reward to a player p_i is:

$$r_i = \frac{\frac{G_i}{N_i} - k_i}{c} \tag{4}$$

where G_i is the total controlling time of the resource, N_i is the number of moves the player has performed so far, k_i is the cost for each move, and c is a constant for normalization. This reward function is designed based on the motivation that it should encourage the player to maximize the ownership of the resource with minimum cost. Overall, the rewards corresponding to different actions are demonstrated in Table II.

TABLE II
REWARDS FOR DIFFERENT ACTIONS TO A PLAYER p_i

Action a_t	r_t
1 (flipping)	$\frac{\frac{G_i}{N_i} - k_i}{c}$
1 (consecutive)	$-k_i$
0 (no-move)	0

B. Model Architecture

We create an LM-adaptive strategy by introducing the Q-learning algorithm to the FlipIt framework, where an agent learns to find an optimal strategy using the LM information of the unknown opponent.

Value Estimation. An agent learns the optimal behaviors to compete against the unknown opponent(s) according to limited feedback from the dynamic environment in real-time. The estimated value of action a_t in state s_t ($Q(s_t,\ a_t)$) is based on the immediate feedback from the environment after each action and the accumulated information during the game. The true value of action a_t in state s_t ($V_{s_t,\ a_t}$) is an immediate reward plus the discounted estimated future rewards which is defined as follows:

$$V_{s_t,\ a_t} = r_t + \gamma\ max_{a_{t+1}} Q(s_{t+1},\ a_{t+1}) \tag{5}$$

where r_t is an immediate reward and γ is a discount factor (a constant in the range of 0 to 1). After each time step, the estimated value $Q(s_t,\ a_t)$ will be updated following the temporal-difference Q-Learning updated rule based on $V_{s_t,\ a_t}$ and the difference between $V_{s_t,\ a_t}$ and $Q(s_t,\ a_t)$. There are various kinds of update rules such as (6), which is applied in our experiments:

$$Q\left(s_t,\ a_t\right) = Q\left(s_t,\ a_t\right) + \alpha\ \left[r_t + \gamma\ max_{a_{t+1}} Q\left(s_{t+1},\ a_{t+1}\right) - Q\left(s_t,\ a_t\right)\right] \tag{6}$$

where α is a learning rate in the range of 0 to 1. The larger the value of α, the less important the old value of $Q(s_t,\ a_t)$.

Action Selection. When selecting an action, a decaying-ϵ greedy exploration strategy can be employed to make a trade-off between the exploitation of the learned $Q(s_t,\ a_t)$ and the exploration of the new action and new state. Specifically, if a randomly generated number is less than ϵ, the agent will choose an action at random, otherwise, he will choose the action that is with the maximum Q value. The decaying-ϵ greedy exploration strategy is represented in (7).

$$a_t = \begin{cases} \text{choose an action at random with probability } \epsilon \\ \text{argmax } Q(s_t,\ a_t) \text{ with probability 1-}\epsilon \end{cases} \tag{7}$$

where ϵ is decaying.

The algorithm of a two-player single-agent FlipIt model is defined in Algorithm 1.

Algorithm 1 Single-Agent RL-based FlipIt Model

```
Initialize ϵ
Initialize state
Initialize 2D Q-table with all zeros
for t ∈ {1, 2, 3, ...} do
    decay ϵ;
    select an action according to (7);
    perform a_t and get the next state s_{t+1}, and reward r_t
    from the environment;
    update Q(s_t, a_t) according to (6);
    s_t=s_{t+1}.
end for when terminal condition is met
```

C. The Uncertainty and Openness of the FlipIt Game

We perform several experiments to study the uncertainty and openness of the FlipIt game based on the the application of the Q-learning algorithm. It is assumed that in all of our experiments, the costs per move for a defender and attacker are 5 and 1, respectively, and the non-adaptive player moves periodically with a default move interval/period of 50 time units.

1) Single-Agent Model: We first define p_0 as a defender playing periodically with a default move period of $p = 50$, and p_1 as an attacker that employs an LM-adaptive strategy interacting with the dynamic environment. We add random noise with specific parameters to the periodic player's moving intervals, and observe how the two players behave.

Adding random Noise. The parameters related to the random noise are $prob$ and $scale$ which indicate the probability of adding a noise and the scale of the noise, respectively. After adding a random noise given a probability, the new move interval becomes $p_\prime$:

$$p_\prime = p + noise \tag{8}$$

where the value of $noise$ is in the range of $[-scale, scale]$, leading to the new move interval $p_\prime$ be in the range of $[p - noise, p + noise]$. Different values of *prob* and *scale* are trialed in the experiments, e.g. $scale \in \{1, 2, 3, ...20\}$, $prob \in \{0.2, 0.4, 0.6\}$.

Swapping the Roles of the Two Players. A simple FlipIt game can be treated as symmetric between two players. However, due to the assumptions that a defender has some priorities over an attacker in this research, the game becomes asymmetric. Therefore, we swap the roles of the two players so that $p0$ becomes the periodic attacker and $p1$ is the LM-adaptive defender, and then evaluate the experiment.

2) Multi-Agent Model: Inspired by MARL, we further extend the model to which both the defender p_0 and attacker p_1 play with LM-adaptive strategies. Each player can get information of the LM of the opponent and selects an action according to (7). The environment updates accordingly based on the joint action, and generates corresponding reward and new state observation to each player. The Q-value for each agent updates according to (6).

Action Combinations and Outcomes. Generally, there are four kinds of action combinations: (0, 1), (1, 0), (0, 0), and (1, 1) (0: no-move, 1: move). Table III explains each action combination and the corresponding cost. Table IV shows the next observation after each action, where k_0 and k_1 are the move costs for p_0 and p_1, respectively, and LM_{opp} is the LM of the opponent.

TABLE III
ACTION COMBINATIONS AND EXPLANATIONS

Action Combination	Explanation	Cost
(1, 0)	p_0 moves, p_1 no move	$(k_0, 0)$
(0, 1)	p_1 moves, p_0 no move	$(0, k_1)$
(1, 1)	p_0 and p_1 both move	(k_0, k_1)
(0, 0)	neither p_0 nor p_1 move	$(0, 0)$

TABLE IV
MOVE TYPES AND OUTCOMES IN THE MULTI-AGENT MODEL

Action a_t	Outcome
1 (flipping)	$s_{t+1} = t - LM_{opp} + 1$
1 (consecutive)	$s_{t+1} = s_t + 1$
0 (no-move)	$s_{t+1} = s_t + 1$

Introducing New Participants to the Model. The above FlipIt game involves only two players, however, in real practice new participants may join in during the game. We perform it by adding new attackers at random time steps, respectively, with a certain probability, and further extend the game from two-player to n-player. The new attackers share the same attributes as the original attacker such as the LM-adaptive strategy and the same move cost, and intend to compete with all the other players to control the resource as long as possible. This means the new attackers not only compete with the defender p_0 but also with the original attacker p_1. In this case, each player can get feedback of the LM of the opponent who is controlling the resource upon performing a move. Furthermore, we vary the parameters related to this experiment including the number of players and the probability of adding a new attacker.

V. Experimental Results

We evaluate a player's strategy by calculating the average benefit using the accumulated ownership of the resource minus the total cost and then divided by the total number of moves that have been performed. We prove that by applying RL in the FlipIt game, an LM-adaptive agent can find an optimal strategy against a variety of opponents including a periodic player (with random noise), an LM-adaptive opponent, and even against multiple adaptive players, and the average benefits for players can always converge to the optimal.

A. Single-Agent Model

We consider first by an LM-adaptive attacker p_1 against a defender p_0 who plays periodically with a default move interval of $p = 50$. According to Fig. 3 , p_1 can find an optimal strategy as the average benefit eventually converges to the optimal against p_0. We also found that the best strategy is to move right after the periodic defender, which is consistent with the conclusion proposed by [7]. With this strategy, the longest ownership of the resource for the agent is 49 time units in each 50 time units as represented in Fig. 4, which greatly outperforms the periodic strategy.

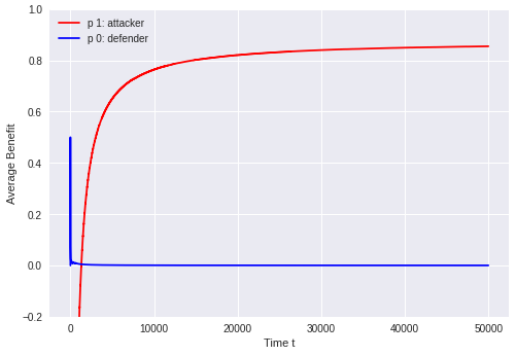


Fig. 3. Average Benefits for Players: Periodic Defender p_0 vs Adaptive Attacker p_1.

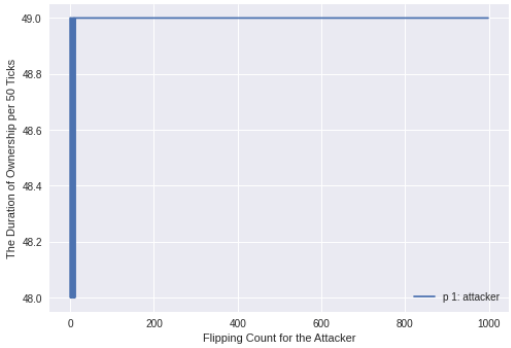


Fig. 4. The Duration of Ownership per Period for the Agent. The "flipping count for the attacker" means the number of times that the agent grabs resource from the opponent during the game.

Given the above, we add random noise to the move periods of p_0 and assign specific values to the parameters $prob$ and $scale$. For example, $prob = 0.2$ and $scale = 5$ mean that the probability of adding a noise is 0.2 and the value of the noise is in the range of [-5, 5]. Fig. 5 shows that in this case, the adaptive player eventually finds an optimal strategy to compete against a periodic opponent even with random noise.

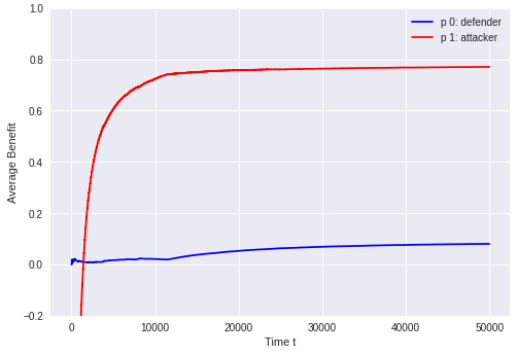


Fig. 5. Average Benefits for Players: Periodic Defender p_0 vs Adaptive Attacker p_1 with random Noise. Here, $prob = 0.2$ and $scale = 5$. Therefore, the period with random noise is in the range of [45, 55].

Furthermore, we evaluate the change of the average benefit for each player by varying the values of $prob$ and $scale$. According to Fig. 6, generally, if keeping a certain $prob$ ($prob \in \{0.2, 0.4, 0.6\}$) and increasing $scale$ from 1 to 20, the average benefit for p_0 decreases and for p_1 it increases. In addition, the larger the value of $prob$, the faster the average benefit changes for each player with the increase of $scale$. Apart from these, the larger the values of the parameters, the longer duration required for convergence.

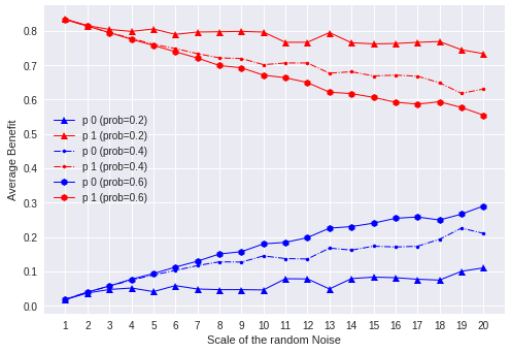


Fig. 6. Average Benefits for Players with the Change of Scale: Periodic Defender p_0 with random Noise vs Adaptive Attacker p_1. The parameters related to the noise: $prob \in \{0.2, 0.4, 0.6\}$ and $scale \in \{1, 2, 3, ...20\}$.

In another experiment, the roles of the defender and attacker are swapped so that p_1 represents the LM-adaptive defender and p_0 is the periodic attacker with random noise, and the parameters related to the noise are: $prob = 0.2$ and $scale = 5$. Fig. 7 shows that the learned strategy for p_1 achieves convergence against p_0, and p_1 can play with a maximum average benefit. Obviously, it is best for p_1 to move at the same time as p_0. This is due to the assumption that if a defender and attacker move together, the defender will gain the

ownership the resource. Therefore, with this optimal strategy, the defender can always control the resource. Finally, we again evaluate the effect of the parameters on the experiment. Given $scale$ being increased from 1 to 20, the larger the $prob$, the greater the fluctuation of the average benefit for each player, which is displayed in Fig. 8.

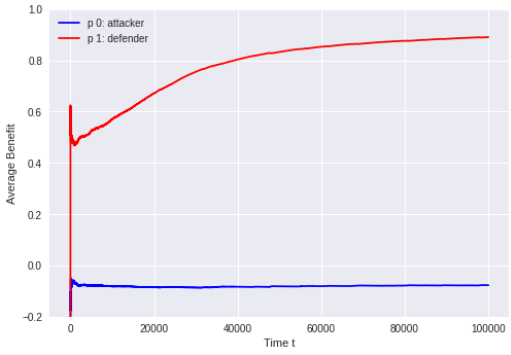


Fig. 7. Average Benefits for Players: Adaptive Defender p_1 vs Periodic Attacker p_0 with random Noise. The parameters related to the noise are: $prob = 0.2$ and $scale = 5$.

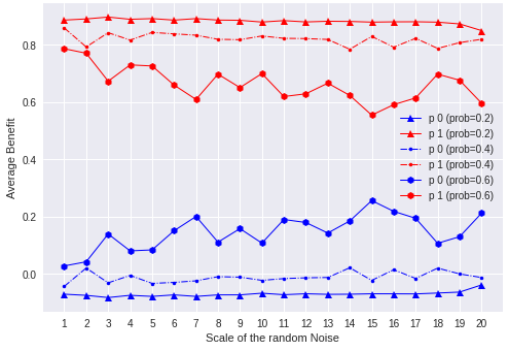


Fig. 8. Average Benefits for Players with the Change of Scale: Adaptive Defender p_1 vs Periodic Attacker p_0. The parameters related to the noise: $prob \in \{0.2, 0.4, 0.6\}$ and $scale \in \{1, 2, 3, ...20\}$.

B. Multi-Agent Model

We let the defender p_0 and the attacker p_1 both play adaptively with the LM information of the opponent. According to Fig. 9, the adaptive strategy learned by either $p0$ or $p1$ converges to the optimal and clearly the defender outperforms the attacker.

Next, we add additional attacker(s) to the previous two-player model at random time steps individually given a specific probability P. We assume that if each attacker is added successfully, the game has up to N players. However, due to the uncertainty, there probably be fewer than N players unless we set $P = 1$. As shown in Fig. 10, we add eight new attackers individually with a certain probability $P = 0.3$ at random time steps causing $N = 10$. Because of $0 < P < 1$, some attackers are not added successfully and the number of players in the end is less than 10. Fig. 10 shows that the LM-adaptive strategy learned by each player gradually converges to the optimal. In addition, we prove that changing the probability P does not affect the convergence but would have influence on the average benefits (we only consider the average benefits for the original two players p_0 and p_1). In general, the larger the probability P, the lower average benefit for p_0 or p_1, which is demonstrated in Fig. 11.

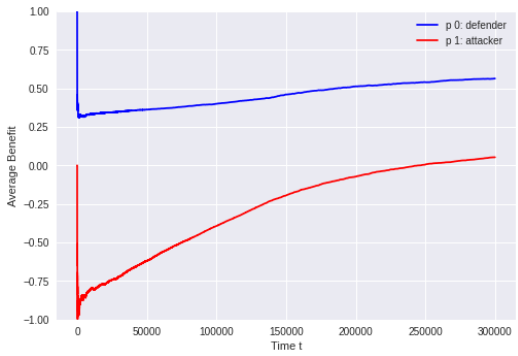


Fig. 9. Average Benefits in a Multi-Agent Model: Adaptive Defender p_0 vs Adaptive Attacker p_1

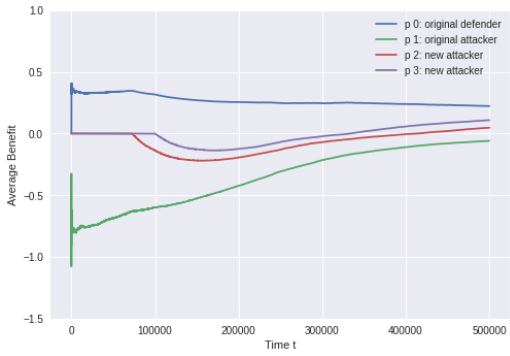


Fig. 10. Average Benefits in a Multi-Agent Model with Additional New Attackers. The parameters in this case: $N = 10$, $P = 0.3$. Due to $P < 1$, the actual number of players is less than 10.

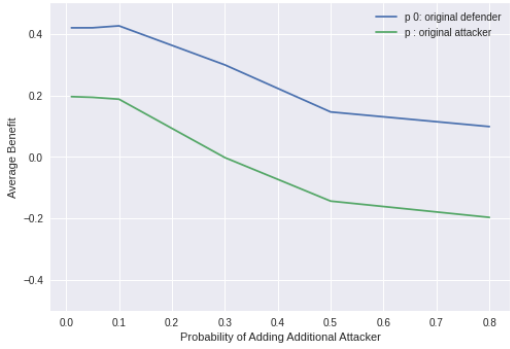


Fig. 11. Average Benefits for p_0 and p_1 with the Change of P

VI. Conclusions and Future Work

This paper considered several variations of the FlipIt game to solve more challenging security scenarios based on the application of temporal-difference Q-Learning. It proved that an LM-adaptive player can find an optimal strategy against a variety of opponents including a periodic opponent (with random noise), an LM-adaptive opponent, and even against multiple adaptive players. Specifically, when an agent competes against a periodic player with random noise, varying the parameters of the noise (i.e. the probability and scale) does not change the result of convergence but affect the average benefits for both players. Furthermore, when both players apply LM-adaptive strategies, each of them can find an optimal strategy and the game eventually achieves a state of Nash equilibrium. This is generalized to an n-player game. Finally, we considered the openness of the game that new participants (LM-adaptive attackers in this case) can be added randomly and individually given a certain probability. Varying the probability does not affect the convergence but influences the average benefits for players, and the larger the probability, the lower the average benefits for the original defender and attacker. Overall, we believe we have confirmed experimentally that Reinforcement Learning can work effectively in the FlipIt security game. Especially, our work can tackle well the noisy information and the openness of the FlipIt game with the application of RL under partial observability.

However, there are still some limitations in this research, e.g. the assumptions of the experiments do not always hold in real-world scenarios, and only the discrete time steps are considered. In future work, we would consider more types of moves and other non-adaptive strategies. We intend on analyzing the multi-objectives problem in the multi-agent FlipIt game and exploring the continuous version of FlipIt with the application of deep learning techniques. We would also like to study the network containing N nodes based on Software-Defined Networking (SDN), rather than only one node (i.e. a single resource) described in the basic FlipIt game. As one of the most promising technologies, SDN makes the network programmable, virtualizable, and equipped with centralized control and a global view [19]. With these advantages, we can more easily and efficiently apply the RL-based model in SDN in the future.

References

[1] Vance, Andrew. "Flow based analysis of Advanced Persistent Threats detecting targeted attacks in cloud computing." 2014 First International Scientific-Practical Conference Problems of Infocommunications Science and Technology. IEEE, 2014.

[2] Fossi, Marc, et al. "Symantec internet security threat report trends for 2010." Volume XVI (2011).

[3] Ghafir, Ibrahim, and Vaclav Prenosil. "Advanced persistent threat attack detection: an overview." Int J Adv Comput Netw Secur 4.4 (2014): 5054.

[4] Bowers, Kevin D., et al. "Defending against the unknown enemy: Applying FlipIt to system security." International Conference on Decision and Game Theory for Security. Springer, Berlin, Heidelberg, 2012.

[5] Roy, Sankardas, et al. "A survey of game theory as applied to network security." 2010 43rd Hawaii International Conference on System Sciences. IEEE, 2010.

[6] Van Dijk, Marten, et al. "FlipIt: The game of "stealthy takeover"." Journal of Cryptology 26.4 (2013): 655-713.

[7] Oakley, Lisa, and Alina Oprea. "QFlip: An Adaptive Reinforcement Learning Strategy for the FlipIt Security Game." International Conference on Decision and Game Theory for Security. Springer, Cham, 2019.

[8] Hamasaki, Koji, and Hitoshi Hohjo. "Moving Target Defense for the CloudControl Game." International Workshop on Security. Springer, Cham, 2021.

[9] Chonka, Ashley, et al. "Cloud security defence to protect cloud computing against HTTP-DoS and XML-DoS attacks." Journal of Network and Computer Applications 34.4 (2011): 1097-1107.

[10] Pham, Viet, and Carlos Cid. "Are we compromised? Modelling security assessment games." International Conference on Decision and Game Theory for Security. Springer, Berlin, Heidelberg, 2012.

[11] Aron Laszka, Gabor Horvath, Mark Felegyhazi, and Levente Buttyán. Flipthem: Modeling targeted attacks with flipit for multiple resources. In Radha Poovendran and Walid Saad, editors, Decision and Game Theory for Security, pages 175–194, Cham, 2014. Springer International Publishing

[12] Farhang, Sadegh, and Jens Grossklags. "FlipLeakage: a game-theoretic approach to protect against stealthy attackers in the presence of information leakage." International Conference on Decision and Game Theory for Security. Springer, Cham, 2016.

[13] Laszka, Aron, Benjamin Johnson, and Jens Grossklags. "Mitigating covert compromises." International Conference on Web and Internet Economics. Springer, Berlin, Heidelberg, 2013.

[14] Greige, Laura, and Peter Chin. "Reinforcement Learning in FlipIt." arXiv preprint arXiv:2002.12909 (2020).

[15] Butt, Umer Ahmed, et al. "A review of machine learning algorithms for cloud computing security." Electronics 9.9 (2020): 1379.

[16] Kang, Chunjian, et al. "An automatic algorithm of identifying vulnerable spots of internet data center power systems based on reinforcement learning." International Journal of Electrical Power Energy Systems 121 (2020): 106145.

[17] Han, Yi, et al. "Reinforcement learning for autonomous defence in software-defined networking." International Conference on Decision and Game Theory for Security. Springer, Cham, 2018.

[18] Alagoz, Oguzhan, et al. "Markov decision processes: a tool for sequential decision making under uncertainty." Medical Decision Making 30.4 (2010): 474-483.

[19] Shin, Myung-Ki, Ki-Hyuk Nam, and Hyoung-Jun Kim. "Software-defined networking (SDN): A reference architecture and open APIs." 2012 International Conference on ICT Convergence (ICTC). IEEE, 2012.

Cyber exclusions: An investigation into the cyber insurance coverage gap

1st Frank Cremer
Kemmy Business School
University of Limerick
Limerick, Ireland
frank.cremer@ul.ie

2nd Barry Sheehan
Kemmy Business School
University of Limerick
Limerick, Ireland
barry.sheehan@ul.ie

3rd Michel Fortmann
Centre for Reinsurance Research
T.H. Köln -
University of Applied Sciences
Cologne, Germany
michael.fortmann@th-koeln.de

4th Martin Mullins
Kemmy Business School
University of Limerick
Limerick, Ireland
martin.mullins@ul.ie

5th Finbarr Murphy
Kemmy Business School
University of Limerick
Limerick, Ireland
finbarr.murphy@ul.ie

Abstract – The importance of cyber insurance as a tool for financial resilience to mitigate the accelerating corporate losses caused by cybercrime is growing. However, there exists a lack of standardization and mutual understanding in cyber insurance policies. With less than a third of cyber insurance claims paid in 2017 in the U.S., there exists a significant gap between the cyber risks businesses need to cover and those actually covered through their cyber insurance policies. This research uses inductive qualitative content analysis to examine the existing exclusions in the terms and conditions of 40 German cyber insurers and compares the summarized results with existing cyber risk events. We posit that the lack of understanding of cyber policy wordings related to cyber risks is a significant problem for companies that could suffer significant losses. The resulting categorization of 15 exclusions and interrelationships with cyber risk events will support businesses, the insurance industry, and researchers in their efforts to understand, measure, and manage cyber risk.

Keywords—cyber insurance, cyber risk, contract design, exclusions, policy wordings, insurance policy

I. Introduction

By 2020, Germany's total amount of losses due to cyber risks in the form of cybercrime (e.g., espionage, theft, and sabotage) doubled to € 223 billion compared to the years 2018 and 2019 [1]. The significance of cyber risks are underlined by the high ranking in the Allianz Risk Barometer 2021 and the announcement of the European Council in April 2021 to establish a center of excellence for cyber security to focus investments on research, technology, and industrial development [2, 3]. Cyber-attacks such as the ransomware attack against oil pipeline operator Colonial Pipeline in 2021 threatened gasoline supplies in parts of the U.S. in the short term [4]. Promptly afterwards, a separate attack on the U.S. I.T. service provider Kaseya disabled the operation of cash register systems, causing a business interruption for up to 1,500 companies worldwide [5].

Although cyber insurance can positively impact the enterprise's cyber resilience, corporate efforts to protect against these risks are relatively low [6]. Measured against the advanced U.S. insurance market with a total premium volume of $1.28 trillion in 2020 and a cyber insurance premium volume of $2.74 billion, the percentage share of cyber insurance in 2020 is less than one per cent [7, 8]. The low willingness to insure against cyber risks could be due to the low acceptance of cyber risks and the lack of understanding of cyber insurance [9]. In 2017, only 28.4 % of cyber insurance claims were paid, with the result that the companies themselves had to pay for losses [10]. This research investigates the efficacy of existing German cyber insurance coverage using an inductive qualitative content analysis of cyber insurance policy exclusions and compares the results with realized cyber risk events. This comparison highlights the coverage gap between cyber insurance policies and corporate cyber exposure.

All precautions must be taken within the risk management framework to avoid, retain, minimize, or transfer cybersecurity risks in a corporate context. Cyber insurance is one way of transferring cyber risks. In general, cyber insurance coverage can be divided into three categories. First-party coverage ensures losses that directly affect the company itself, such as business income loss, costs for data and system restoration [11]. Third-party coverage extends to third-party liability claims. The insurance cover is, for example, intellectual property and media breaches, cost resulting from legal proceedings and fines [12]. Assistance services are often listed as the third category. This includes the services provided by the insurer or a company contracted by it in the event of a cyber incident. These include, for example, security auditing, incident response teams, public relations and digital forensics expenses [13]. Cyber insurers can positively influence cyber security measures at companies, acting as proxy regulators by requiring certain minimum levels of security from companies for the insurance cover they provide [14, 15].

Companies that already have cyber insurance coverage face issues in understanding it accurately. The reasons underlying this understanding deficit includes the lack of experience in their treatment of cyber risks, no standardized definitions, and

overlapping insurance coverage with other insurance coverage lines [16]. As a result, the cyber insurance terms and conditions are imprecise, and the contractually agreed coverage is inadequately described [17]. In addition, the information on insured cyber damage is unclear. However, the wording of insurance terms and descriptions of contracted coverage is the structure of cyber coverage [18]. For stakeholders of cyber insurers, the problem arises that they do not know exactly which loss scenarios are covered and which are specifically excluded [17]. While cyber insurers indicate their exclusions, these are not standardized and vary widely, creating coverage gaps for insured companies. This makes it necessary to systematically identify, analyze and comprehensively quantify the existing exclusions of cyber insurance general terms and conditions [19].

This research uses inductive qualitative content analysis to examine the existing exclusions in the terms and conditions of German cyber insurers and compares the summarized results with existing cyber risk events. The cyber insurers under review represent a combined market share of over 90% in Germany. The study provides a comparative analysis of the associations between exclusions in cyber policy wordings and cyber risk events. There are few studies that consider the provider side of cyber insurance[12, 20]. Some studies have looked at inclusions and exclusions but have not related them to cyber risk events [11, 21]. This paper is the first research to close the gap with its contribution through a comprehensive analysis of the exclusions in the general terms and conditions of the cyber insurers operating in Germany.

For researchers involved in cyber risks and cyber insurance, a comprehensive overview of exclusions used by German cyber insurers in their general terms and conditions is provided. The research results can be used to identify differences in wordings between countries and large and small cyber insurers. In addition, the results can provide evidence for more comprehensive investigation into individual exclusions. For example, deeper analysis of exclusion clause war has already been completed [22, 23], but there hasn't been further research into silent exposures emanating from critical infrastructure outages or deliberate defamation exclusions. For policymakers, the results show an overview of which cyber losses are not covered in the general terms and conditions. The research paper could provide guidance on the extent to which the state needs to be active and in which areas to ensure comprehensive cyber coverage.

The remainder of the paper is structured as follows. Section two describes the related work regarding cyber insurance. Section three outlines the mixed research method used in this work and the process. Section four details the results of the identified exclusions. Further discussion is included in section five before concluding the article in section six.

II. Related work

Due to the high significance of cyber risks and their risk transfer, some research has already been conducted in this field. In their pioneering survey, Marotta et al. [19] summarized the current state of knowledge from practice and science on cyber insurance. The authors examine 14 cyber insurance policies and propose a list of possible research directions related to cyber insurance. One of these research areas include exploring the inclusions and exclusions of cyber insurance terms and conditions, which other researchers also indicated in their further research [20].

From a risk management perspective, Majuca et al. [24] analyze one of the first cyber insurance products. The authors consider the different coverages of cyber insurance and analyze the inclusions and the exclusions of cyber insurance. In this research, seven different wordings of cyber insurers are considered. The researchers conclude that it may take a significant loss event to increase the market penetration of cyber insurance. Another consideration of the provider side of cyber insurance takes the research qualitative research on the provider side of cyber insurance was conducted by Romansky et al. [11], which analyzed the underwriting process for cyber insurance and revealed how cyber insurers understand and assess cyber risks. For this research, the researcher examined 235 American cyber insurance policies that were publicly available and looked at three components (coverage, application questionnaires and pricing). In the findings, the authors note that many of the insurers used simple flat rates (based on a single calculation of expected loss), while others included more parameters such as the company's asset value (or company revenue) or standard insurance metrics (e.g., deductibles, limits) and industry in the calculation. Regarding the insurance coverages, the authors have been able to identify the common inclusions as well as exclusions.

Further research on exclusions in cyber insurance was conducted by Ferland [23]. Specifically, this paper takes an in-depth look at the exclusion war clause. In the context of insurance litigation (Zurich vs Mondolez), the paper analyses the attribution of a cyber-attack to a country, and the interpretation of the war exclusion in an insurance policy in the cyber context. The extent to which the war exclusion clause can be interpreted and construed in the context of insurance law is discussed. Further research on the war exclusion clause is presented from Woods and Weinkle [22]. The paper examines war clauses in 56 cyber policies through inductive qualitative content analysis. In their analysis, the researchers concluded that, due to market and regulatory forces, insurers specialized in cyber insurance have arrived at a balance that excludes circumstances well below the threshold for war but includes cyber-terrorism. The research by Wrede et al. [16] examines the design of affirmative and silent coverage regarding cyber risks in traditional insurance policies for selected product lines on the German market. The authors use for research an approach from an inductive qualitative content analysis of the general terms and conditions of cyber insurance policies and structured interviews with experts from the German insurance industry. Overall, the authors' results show a partial vulnerability in different insurance lines due to the current design of insurance terms and conditions.

This paper is the first to comprehensively analyze, categorize, and summarize exclusions in general terms and conditions of 40 cyber insurers in Germany. It further contributes to the current understanding by providing an overview of the exclusion categories and their frequencies in the policy

wordings. Finally, this work compares the exclusions to current cyber risk events and identifies risk areas and interrelationships not currently covered by general terms and conditions.

III. Methodology

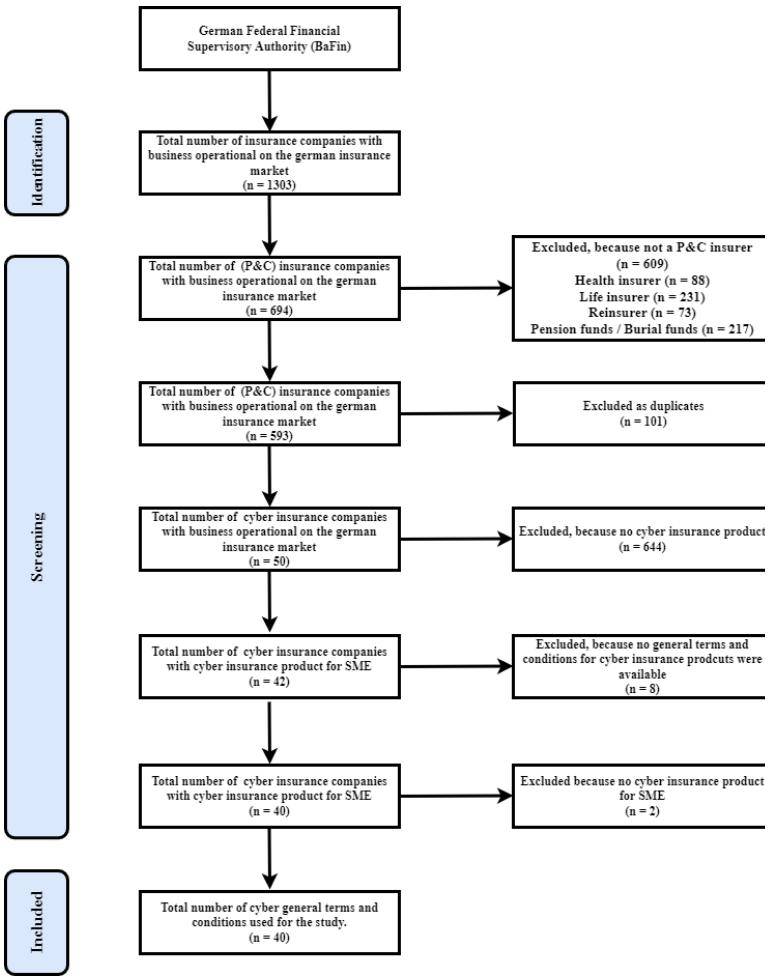
A database of 40 cyber insurance policy terms and conditions was created using publicly available policy information (available online or by request). The 40 cyber insurers, which are representative of approximately 90% combined market share in Germany, were selected by criteria outlined in the following section and illustrated in Figure 1.

This research focuses on cyber insurers that are active in the German market, whose cyber insurance serves small and medium-sized corporate customers, and who fall under the German Federal Financial Supervisory Authority (BaFin). The information on the selected cyber insurance companies was obtained from BaFin, which provides databases on the insurers it supervises.

First, entries from BaFin's company databases were transferred into an Excel file. The file generated from BaFin's company database initially contained a total of 1303 datasets. The individual data sets contained information on the names of the insurance companies' business areas, as well as the country of origin of the company's registered office and the corresponding address. Based on the framework of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [25], the targeted cyber insurers were identified through a multistep process outlined in Figure 1.

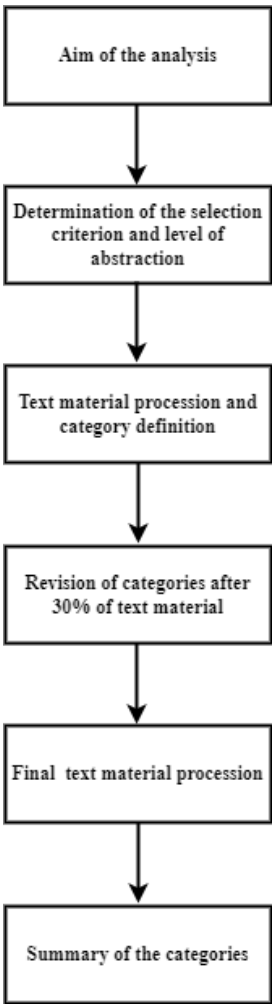
All insurers whose business was not Property and Casualty (P&C) were removed from the analysis in the first step. According to this criterion, 609 insurers were excluded. After that, possible duplicates in the form of subsidiaries or similar were sorted out. This left 593 insurers for further filtering after 101 duplicates were excluded. The websites of the remaining insurers were manually checked to determine whether cyber insurance was offered. The criterion was fulfilled if this information was visible. As a result, 50 cyber insurers in Germany could be identified. The collection of the terms and conditions of the policies to be examined was carried out in December 2021 and January 2022. For this purpose, the publicly accessible general terms and conditions of the identified cyber insurers were collected via their websites. If these possibilities did not exist, the documents were requested via email from the respective insurer. In this way, 42 general terms and conditions could be determined in the scope of the investigation. The information was reviewed by a panel of experts whose members have previously published on cyber liability, cyber risk and cybersecurity topics in peer-reviewed journals. In addition, the expert group members have in-depth knowledge of insurance-related cyber risk topics. This includes, for example, the legal consideration of insurance wordings. The cyber policy wordings were only included in the analysis if they cover small and medium enterprises (SME). Finally, 40 general cyber insurance terms and conditions could be analyzed within the scope of this research paper.

Figure 1: Flow diagram of the selection process



In the further step, this study methodically relies on a systematic evaluation of text documents. The development of an overview of categories based on the research objectives and the coding of the available text material using this differentiated category scheme characterizes a qualitative content analysis. For the research question of this study, the general terms and conditions of insurance of the identified cyber insurers that were available for this study were analyzed. The general terms and conditions of insurance products provide a suitable data basis for the research, as they enable the description of the scope of insurance coverage [18]. In addition to other contractual components, these also include exclusions, which exclude certain categories and scenarios from insurance coverage. The general insurance terms and conditions of the insurance products to be examined can be divided into standalone cyber insurance and bundled cyber insurance for SMEs. The reason for this selection is the easier availability of policy wording as well as the uniform structure, as industrial risks often have highly individualized clauses.

Figure 2: Sequence of inductive category formation adapted from Mayring [26]



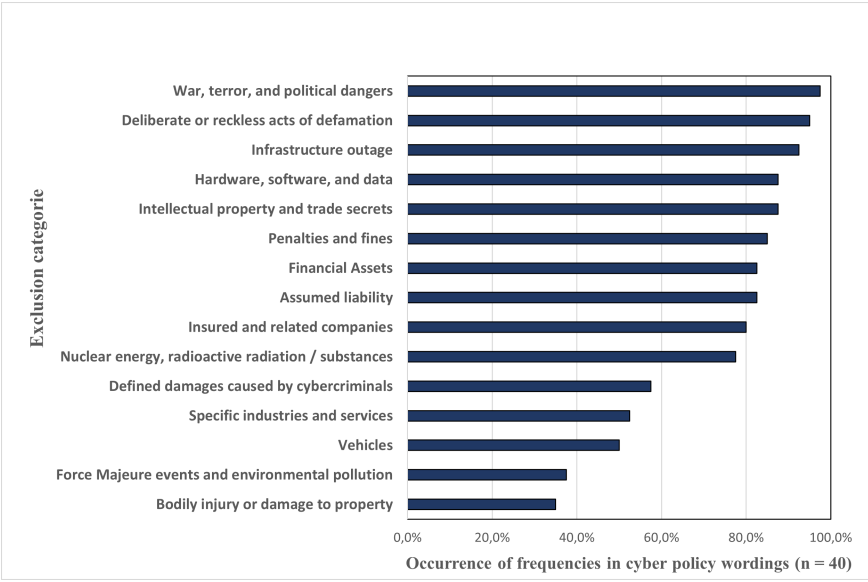
The evaluation was based on Maying's model of qualitative content analysis, which enables a rule-based and systematic evaluation of the data [26]. The qualitative data analysis software MAXQDA was used to facilitate and systematically structure the coding process. Due to the nature of the text material at hand, a pragmatic and appropriate combination of the two analysis techniques, inductive category formation and content-based structuring, was carried out [16, 26]. Since the categories were derived directly from the conditions of assurance to be analyzed, the method is particularly suitable for the study of phenomena from practice as well as for problems whose academic knowledge is limited in terms of literature and research [27].

An inductive qualitative content analysis approach was followed, in which individual categories are derived directly from the material without being inferred based on existing theoretical concepts [28]. This approach aims directly at drawing conclusions to answer the research objective and is applied to the 40 general insurance terms and conditions on cyber insurance. In the first step, the complexity of the qualitative data material was reduced successively and manually by coding the documents. Here, the data material was checked for relevant text segments, which were marked and assigned "text-related" codes. Relevant text segments were those marked as "exclusion," "uninsured perils," and "deemed uninsured" in the policy wordings.

The text segments obtained through this process form the basis for inductive category formation. The entire text material was first to read several times to code it in successive degrees of abstraction. In the next step, preliminary category formulations were developed from the material through several iterations, which were applied as descriptive code. The text segments were then structured using the descriptive code and transformed into preliminary categories. After approximately 30% of the material was collected and coded, a revision of the categories was made for robustness. The research team reviewed the previous data material, the codes and the provisional categories and formulated applicable coding rules to ensure the quality of the coding [29]. This was followed by the coding of the remaining general insurance conditions.

The coded text material was summarized and evaluated regarding preliminary categories. If the categories were often represented identically in terms of language in the cyber policy wordings, a main summary category was formed. This was necessary to determine the scope for each category and to form subcategories on that basis to ensure that the different expressions of each category were well defined. The final category scheme contained a total of 15 categories and was obtained from 40 cyber policy wordings.

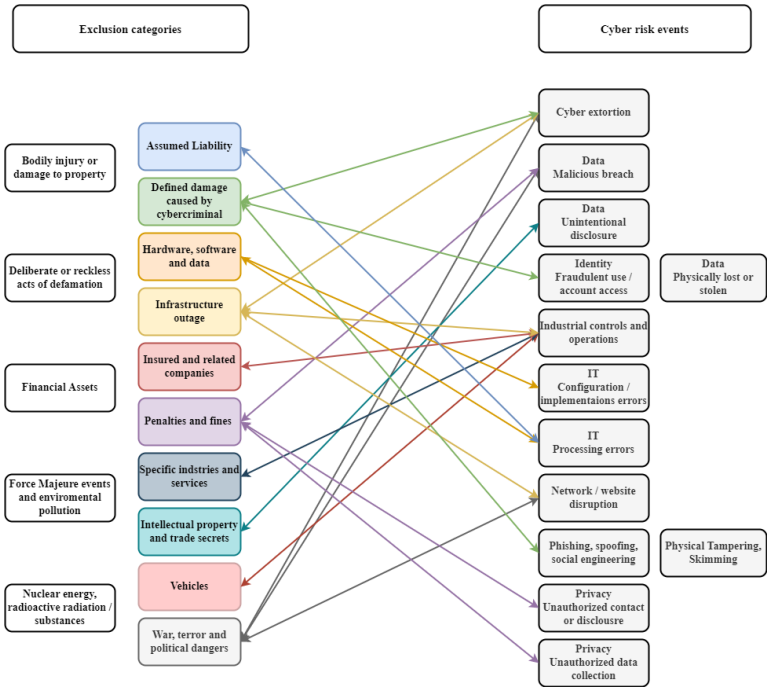
Figure 3: Name of exclusions and percentage of occurrence in the considered wordings



For the second part of the results, an overview was compiled to obtain information on which exclusion categories are directly affected by cyber risk events and could therefore be of particular importance for companies. For this purpose, the

expert panel examined the respective exclusion categories and the interrelationships between cyber risk events. The cyber risk events used were drawn from the Advisen database, primarily from the cyber loss data collection [30]. The cyber risk events had already been used in other research and were suitable for classification due to their respective descriptions [31, 32]. The descriptions of the cyber risk events can be found in the appendix. In the review, the expert panel determined which cyber risk categories could be assigned to the individual exclusions. As the mapping only considered direct cyber risk scenarios for the respective categories described, it was not possible to allocate all exclusion categories to cyber risk events. The next Figure shows which exclusion category was assigned to the respective cyber risk events.

Figure 4: Connections between exclusions and cyber risk events



IV. Results

The results are divided into two parts. The first part is the results of the qualitative content analysis, which aims to form and summarize inductive exclusion categories from the text segments of the cyber policy wordings. The second part is the subsequent comparative analysis of the exclusion categories with the current cyber risk events.

A. Results of the inductive qualitative content analysis

In this subsection, the results of the qualitative content analysis are highlighted. As a result, the exclusion categories and their summary description are discussed.

Assumed liability
No coverage is provided for any claims, liabilities, losses, or defence costs arising directly or indirectly out of any contractual or assumed liability, warranty or guarantee unless the insured would be legally liable in any event but for such contractual or other assumed liability, warranty, or guarantee. Also included in this category are exclusions in connection with the recall of proprietary or third-party products.

Bodily injury or damage to property
No insurance coverage for the cyber policy wordings under consideration applies to bodily injury or property damage. By way of example, for bodily injury, these include, but are not limited to, physical injury, illness, or death of a person. For property damage, the exclusion applies to the destruction of tangible property, including loss of use thereof.

Defined damage caused by cybercriminals
In the context of cybercrimes, there are various specific exclusions defined by certain insurers. Mostly, these relate to insurance claims resulting from the payment of ransom or extortion, or the fulfilment of extortion demands. Another exclusion relates to damage caused by identity theft or phishing. In addition, one cyber insurer was found to exclude from its insurance covers losses resulting from targeted data manipulation (e.g., fake president fraud).

Deliberate or reckless acts of defamation
Excluded are any claims, liabilities, losses or defence costs arising directly or indirectly from a defamatory statement made intentionally or recklessly by the Insured. This also includes damage caused by a knowing deviation from the law, regulation, resolution, power of attorney or instruction, or by any other knowing breach of duty by the policyholder or his representatives.

Financial Assets
This class combines several exclusions relating to financial assets. These include insurance claims or losses arising from or in connection with any form of purchase or sale of securities, commodities, derivatives, foreign exchange, bonds and similar investments. Furthermore, the outflow of assets has been excluded if these have arisen in the context of a security breach.

Force Majeure events and environmental pollution
In this class, the damage is excluded to the extent that elementary natural forces have had an effect, such as fire, flood, storm, lightning, frost, explosion or extreme weather or temperature conditions. In addition, environmental pollution that contaminates or otherwise adversely affects air quality, the atmosphere, water quality, soils or subsoils, fauna, flora, human health is excluded.

Hardware, software, and data
In this category, there are aspects of information technology that are not insured. These include losses related to planned processes such as maintenance and shutdown of information processing systems, deletion and modification of electronic data, the introduction of new systems or software. Also included in the exclusions are the introduction of new systems and software, untested systems and software or systems and software not yet approved for the intended use. Coverage also does not include costs incurred to correct software errors or security vulnerabilities. In this exclusion class, costs that represent an improvement for the

policyholder are not reimbursed. In addition, the costs associated with the wear and tear of systems, software and data are excluded.

Infrastructure outage
This exclusion, at 91.7%, is one of the most frequent exclusions found in the policy wordings. Excluded are infrastructure outages due to or resulting from electrical or mechanical failures or interruptions, electrical faults, surges, spikes, brownouts, power outages, or failures of electricity, gas, water, telecommunications, or other infrastructure. This also includes the internet and communication via satellite.

Insured and related companies
In these categories, claims are excluded if they involve companies that have a closer connection to the insured company. In many of the coded segments, the exclusions mainly relate to affiliated companies that are capitalistically linked to the policyholder or its shareholders through a shareholding of more than 25 % or are under uniform entrepreneurial management and use the same information and communications technology infrastructure. This also includes claims by relatives of the policyholder, legal representatives, shareholders, and other representatives.

Nuclear energy, radioactive radiation / substances
In many other lines of insurance, nuclear risks are excluded. The same applies to cyber insurance. In this context, insured events or damage caused by nuclear energy, nuclear radiation or radioactive substances are not considered insured.

Penalties and fines
Losses are not considered insured if they violate applicable laws. These include, among others, insured events arising from official enforcement or orders, penalties, contractual penalties, fines, punitive and exemplary damages against the policyholder. In addition, many general terms and conditions of insurance have identified the exclusion of discrimination, with the note Damages due to violation of a provision protecting against discrimination from the General Equal Treatment Act [33].

Specific industries and services
These exclusions target some industries and sectors that pose a much greater risk to cyber insurers in the course of their operations. These include various sectors whose business almost invariably relies on the use of, applies, develops, maintains, and operates information technology. Other sectors include gambling, operators of pornographic content, and certain large-scale risks (e.g., offshore facilities).

Intellectual property and trade secrets
Insurance claims or damages due to or in connection with plagiarism or infringements of patents, trademark rights, copyrights and other forms of intellectual property are excluded from insurance coverage. This also applies to licenses or license fees as well as infringements of competition and antitrust law.

Vehicles
Within the scope of the exclusions for vehicles, it could be determined that insured events or damage in connection with vehicles of any kind are excluded. Examples include motor vehicles, aircraft, rail vehicles, watercraft and spacecraft.

War, terror, and political dangers
The focus in this category is on the exclusion due to war and terror. Damages caused by the clause War and terror are considered excluded in this clause. Examples include war, invasion, insurrection, revolution, or other forms of seizure of power or state-initiated acts (e.g., espionage or cyber warfare). Terror exclusion includes any act intended to achieve political, religious, ethnic, or ideological goals that are likely to spread fear or terror among the population or segments of the population in order to influence a government. This category also includes the exclusion of political dangers. This excludes damage in connection with, among other things, confiscation, expropriation, or destruction of property by a government.

B. Results of the comparative analysis

In the second subsection, the links between cyber risk events and exclusion categories are discussed. This is to illustrate that a certain number of exclusions are exposed to a larger number of cyber risk events. The results are supplemented with brief descriptions.

The second part of the results looks at the correlations between the exclusion categories and the cyber risk events. While 50% of the exclusion categories are only assigned to a single cyber risk event, the remaining half have more than one connection to cyber risk events. The exclusion of the “defined losses caused by cybercriminals” category is linked to 23.1% of potential cyber risk events. For example, cyber extortion is currently a popular means used by cybercriminals to extort money from companies in various ways. In the form of ransomware attacks, they infect companies' software and lockout users. For companies, this represents a significant financial risk. On the one hand, they must pay the ransom demands, and on the other, they are blackmailed into publishing the captured data on the darknet in the event of non-payment. Thus, such an exclusion poses a significant risk for a company that does not fully understand its cyber insurance coverage. In terms of hardware, software and data exclusions are associated with 18.2% of potential cyber events.

The increasing number of companies working with cloud service providers creates a certain dependency. Many companies use cloud platforms to outsource parts of their data storage, processing, analysis, and I.T. infrastructure. As a result, companies are dependent on the services of cloud service providers to function continuously. The exclusion of infrastructure outages could be a major potential loss for many companies if they cannot access the infrastructure they need. At 23.1%, this exclusion also has a high association with other cyber risk events. Cyber risk events that have a connection to infrastructure outages include cyber extortion, industrial controls and operation, and network/website disruption. This exclusion is one of the most common exclusion categories and has a higher probability of occurring due to the higher number of causes. Regarding industrial controls and operation, vehicles should also be included. Due

to the progressive development of connected and autonomous vehicles, these represent an increasing risk potential [34].

Another exclusion category that has different connections to cyber risks is war, terror, and political dangers category. As with the previous exclusion categories, the cyber risk cyber extortion also belongs to this category. The reason for this is the possibility that state-funded cybercriminals or terrorist-motivated cybercriminals could be behind such an attack. Due to the difficulty of detection, cyber insurers could refer to this clause and reject the claims. In addition, sovereign intervention by the government or local authorities on the company could also result in claims not being insured.

V. Discussion

This paper presents an inductive qualitative content analysis on exclusions of 40 general cyber insurance terms and conditions for small and medium-sized enterprises in Germany. In this framework, various exclusion categories were fully identified and then classified and summarized into 15 different categories. The exclusion categories and their associations with cyber risk were then analyzed. The results provided a comprehensive overview of the exclusions in the cyber policy wordings and their connections to cyber risks. Each exclusion category contains a description of which exclusions the category refers to and which examples are available. In this way, the different categories can be compared and related to cyber risks. Of course, this research has its limitations, so our selection of identified exclusions cannot necessarily be considered as a representation of all available cyber insurance terms related to cyber insurance. For example, we considered only the general insurance terms and conditions of German cyber insurers. The wording of the insurance policies included cyber insurance for SMEs as standalone and as bundled insurance. The additional options, in which cyber insurers include some exclusions for an additional premium, were not considered. In addition, only the exclusions were considered; individual areas such as definitions, which can be interpreted as a scope in the interpretation of the insurance coverage, were not addressed. Finally, it should be noted that this is an overview of currently available cyber insurance exclusions, which are subject to constant change.

The results of the exclusion categories show that some exclusions represent a large loss potential for policyholders despite cyber insurance. For this reason, clear and consistent communication of cyber risks and cyber insurance is important [9]. The global cyber insurance market is estimated to be $5.5 billion in 2020 [35]. Compared to the just under $1 trillion in global losses from cybercrime [36], action needs to be taken for both the insurance industry and international commerce to provide complete and understandable insurance coverage as well as appropriate safeguards to minimize cyber risk.

The European Union Agency for cybersecurity clearly shows that cybercriminals represent a major potential threat to companies [3]. For this reason, it is important that companies have comprehensive cyber insurance coverage as part of risk management, but also in terms of financial resilience, and are informed about their possible exclusions.

The results of this research can be used by companies to review their existing insurance coverage. In addition, the correlations between exclusions and cyber risk areas where companies need to take greater action are becoming apparent. The exclusion categories listed make it possible to systematically review the individual areas for the company and initiate appropriate measures. These range from increased security measures to risk management. The presentation of exclusion categories and cyber risks can be used by cyber insurers as an opportunity to review their own cyber insurance products. By providing appropriate insurance supplements that cover the exclusions, new market potential can be tapped. In addition, cyber insurers can become the results to develop a better understanding of the coverage needs of businesses. Many exclusions are not clearly worded and offer the risk that customers will interpret them differently. In addition, this research supports the standardization of insurance terms and conditions. During the research, it was found that many cyber policy wordings varied widely, and there were rarely consistent definitions.

VI. Conclusion

In this paper, we conducted an inductive qualitative content analysis of exclusions in general cyber insurance terms and conditions. We found that some exclusions represent significant and, potentially unperceived, cyber loss exposure for companies. Due to the dynamic nature and lack of historical data, assessing and understanding cyber risks and their risk transfer is a major challenge for all cyber insurance stakeholders. To address this challenge, appropriate measures must be taken to reduce the occurrence of cyber risks, formulate understandable and clearly identifiable exclusions, and highlight their impact.

Companies could consolidate this new knowledge into their corporate culture to address the risk of exclusions in cyber insurance and their connections to cyber risks in more detail. For cyber insurance companies, the exclusions show that many exclusions pose a serious risk to companies and that there is no comparability between different cyber insurance policies due to a lack of standardization of cyber policy wordings. As a result, companies may be in the dark about their coverage. Cyber insurers can use the findings to revise their cyber insurance products. By making appropriate additions to cyber insurance coverage, only a larger premium can be collected, thus contributing to customer satisfaction by allowing customers to respond appropriately by identifying potential gaps in coverage.

This paper proposes several research directions. Currently, there is only limited research on the provider side of cyber insurance. Some topics could have been further researched for this purpose. Examples of areas include pricing, cyber insurance claims processing applications, and cyber policy wording. The last category has already been explored somewhat in this paper with respect to exclusions. Further areas of research may include the study of individual

exclusions, inclusions, coverage extensions, limitations, or the wording of cyber policies in other countries. By compiling different areas of cyber insurance research, the insurance industry can be assisted in publishing consistent policy wordings for policyholders, which will lead to better market penetration [37]. This would benefit not only cyber insurers but also customers, as standardized clauses could lead to better and more understandable insurance coverage [38].

REFERENCES

[1] Bitkom. "German businesses under attack: losses of more than 220 billion euros per year," 11 January, 2022; https://www.bitkom.org/EN/List-and-detailpages/Press/German-business-losses-more-than-220-billion-euros-per-year.

[2] Allianz. "Allianz Risk Barometer," 15 May 2021; https://www.agcs.allianz.com/content/dam/onemarketing/agcs/agcs/reports/Allianz-Risk-Barometer-2021.pdf.

[3] E. Council. "Cybersecurity: how the EU tackles cyber threats," 10 May, 2021; https://www.consilium.europa.eu/en/policies/cybersecurity/.

[4] D. Brower, and M. McCormick, "Colonial pipeline resumes operations following ransomware attack," *Financial Times*, 2021.

[5] J. Panettieri. "Kaseya REvil Ransomware Cyberattack: Hacker Charged," 10 January, 2022; https://www.msspalert.com/cybersecurity-breaches-and-attacks/kaseya-rmm-cyberattack-warning/.

[6] K. Hausken, "Cyber resilience in firms, organizations and societies," *Internet of Things,* vol. 11, pp. 100204, 2020.

[7] I. I. Institute. "Facts + Statistics: Industry overview," 5 September, 2021; https://www.iii.org/fact-statistic/facts-statistics-industry-overview.

[8] FitchRatings. "Sharply Rising Cyber Insurance Claims Signal Further Risk Challenges," 5 August, 2021; https://www.fitchratings.com/research/insurance/sharply-rising-cyber-insurance-claims-signal-further-risk-challenges-15-04-2021.

[9] OECD. "Encouraging Clarity in Cyber Insurance Coverage," 2 January, 2022; https://www.oecd.org/finance/insurance/Encouraging-Clarity-in-Cyber-Insurance-Coverage.pdf.

[10] AMBest. "Market Segment Report: Cyber Insurance Market Sees Steady Growth but Still Awaiting a Real Growth Spurt," 06.12., 2021; https://www3.ambest.com/ambv/sales/bwpurchase.aspx?record_code=273764&altsrc=.

[11] S. Romanosky, L. Ablon, A. Kuehn, and T. Jones, "Content analysis of cyber insurance policies: How do carriers price cyber risk?," *Journal of cybersecurity (Oxford),* vol. 5, no. 1, 2019.

[12] C. Biener, M. Eling, and J. H. Wirfs, "Insurability of cyber risk: An empirical analysis," *The Geneva Papers on Risk and Insurance-Issues and Practice,* vol. 40, no. 1, pp. 131-158, 2015.

[13] G. Peters, P. V. Shevchenko, and R. Cohen, "Understanding cyber-risk and cyber-insurance," *Macquarie University Faculty of Business & Economics Research Paper*, 2018.

[14] J. MacColl, J. R. Nurse, and J. Sullivan, "Cyber insurance and the cyber security challenge," *RUSI Occasional Paper*, 2021.

[15] G. E. Marchant, and Y. A. Stevens, "Resilience: A New Tool in the Risk Governance Toolbox for Emerging Technologies," *UCDL Rev.,* vol. 51, pp. 233, 2017.

[16] D. Wrede, T. Stegen, and J.-M. G. von der Schulenburg, "Affirmative and silent cyber coverage in traditional insurance policies: qualitative content analysis of selected insurance products from the German insurance market," *The Geneva Papers on Risk and Insurance-Issues and Practice,* vol. 45, no. 4, pp. 657-689, 2020.

[17] N. Hare-Brown, "Confusing terminology stunts the growth of cyber insurance," *Computer Fraud & Security,* vol. 2019, no. 4, pp. 16-17, 2019.

[18] D. Woods, and A. Simpson, "Policy measures and cyber insurance: A framework," *Journal of Cyber Policy,* vol. 2, no. 2, pp. 209-226, 2017.

[19] A. Marotta, F. Martinelli, S. Nanni, A. Orlando, and A. Yautsiukhin, "Cyber-insurance survey," *Computer Science Review,* vol. 24, pp. 35-61, 2017.

[20] G. Falco, M. Eling, D. Jablanski, V. Miller, L. A. Gordon, S. S. Wang, J. Schmit, R. Thomas, M. Elvedi, and T. Maillart, "A research agenda for cyber risk and cyber insurance."

[21] J. Bateman, *War, Terrorism, and Catastrophe in Cyber Insurance: Understanding and Reforming Exclusions*: Carnegie Endowment for International Peace., 2020.

[22] D. W. Woods, and J. Weinkle, "Insurance definitions of cyber war," *The Geneva Papers on Risk and Insurance-Issues and Practice,* vol. 45, no. 4, pp. 639-656, 2020.

[23] J. Ferland, "Cyber insurance–What coverage in case of an alleged act of War? Questions raised by the Mondelez v. Zurich case," *Computer Law & Security Review,* vol. 35, no. 4, pp. 369-376, 2019.

[24] R. P. Majuca, W. Yurcik, and J. P. Kesan, "The evolution of cyberinsurance," *arXiv preprint cs/0601020*, 2006.

[25] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The PRISMA 2020

statement: an updated guideline for reporting systematic reviews," *Systematic Reviews,* vol. 10, no. 1, pp. 89, 2021/03/29, 2021.
[26] P. Mayring, "Qualitative content analysis: Theoretical background and procedures," *Approaches to qualitative research in mathematics education*, pp. 365-380: Springer, 2015.
[27] U. Kuckartz, "Qualitative Inhaltsanalyse," 2012.
[28] J. Y. Cho, and E.-H. Lee, "Reducing confusion about grounded theory and qualitative content analysis: Similarities and differences," *Qualitative Report,* vol. 19, no. 32, 2014.
[29] L. Burla, B. Knierim, J. Barth, K. Liewald, M. Duetz, and T. Abel, "From text to codings: intercoder reliability assessment in qualitative content analysis," *Nursing research,* vol. 57, no. 2, pp. 113-117, 2008.
[30] P. V. Shevchenko, J. Jang, M. Malavasi, G. W. Peters, G. Sofronov, and S. Trück, "Quantification of Cyber Risk–Risk Categories and Business Sectors," *Available at SSRN 3858608*, 2021.
[31] K. Palsson, S. Gudmundsson, and S. Shetty, "Analysis of the impact of cyber events for cyber insurance," *The Geneva Papers on Risk and Insurance-Issues and Practice,* vol. 45, pp. 564-579, 2020.
[32] I. Aldasoro, L. Gambacorta, P. Giudici, and T. Leach, "The drivers of cyber risk," 2020.
[33] F. A.-D. Agency. "Act Implementing European Dirctives Putting Into Effect the Principle of Equal Treatment," 02.01., 2022; http://www.ilo.org/dyn/natlex/docs/ELECTRONIC/77201/93329/F695840346/DEU77201.pdf.
[34] B. Sheehan, F. Murphy, M. Mullins, and C. Ryan, "Connected and autonomous vehicles: A cyber-risk classification framework," *Transportation research part A: policy and practice,* vol. 124, pp. 523-536, 2019.
[35] B. Dyson. "COVID-19 crisis could be 'watershed' for cyber insurance, says Swiss Re exec," 7 May, 2021; https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/covid-19-crisis-could-be-watershed-for-cyber-insurance-says-swiss-re-exec-59197154.
[36] Z. Maleks Smith, E. Lostri, and J. A. Lewis. "The Hidden Costs of Cybercrime," 16 May, 2021; https://www.mcafee.com/enterprise/en-us/assets/reports/rp-hidden-costs-of-cybercrime.pdf.
[37] F. Cremer, B. Sheehan, M. Fortmann, K. Arash N., M. Mullins, F. Murphy, and S. Materne, "Cyber risk and cybersecurity: a systematic review of data availability," *The Geneva Papers on Risk and Insurance - Issues and Practice*, 2022.
[38] B. Sheehan, F. Murphy, A. N. Kia, and R. Kiely, "A quantitative bow-tie cyber risk classification and assessment framework," *Journal of Risk Research*, pp. 1-20, 2021.

APPENDIX

With reference to Shevchenko et al. (2021), the cyber risk events of the Advisen "Loss Cyber Data" database are described as follows:

Privacy – Unauthorized Contact or Disclosure: cases when personal information is used in an unauthorized manner to contact or publicize information regarding an individual or an organization without their explicit permission.

Privacy – Unauthorized Data Collection: cases where information about the users of electronic services, such as social media, phones, websites, and similar is captured and stored without their knowledge or consent, or where prohibited information may have been collected with or without their consent.

Data – Physically Lost or Stolen: situations where personal confidential information or digital assets have been stored on, or may have been stored on, computer, peripheral equipment, data storage, or printouts which has been lost or stolen, or improperly disposed of.

Data – Malicious Breach: situations where personal confidential information or digital assets either have been or may have been exposed or stolen, by unauthorized internal or external actors whose intent appears to have been the acquisition of such information.

Data – Unintentional Disclosure: situations where personal confidential information or digital assets have either been exposed, or may have been exposed, to unauthorized viewers due to an unintentional or inadvertent accident or error.

Identity – Fraudulent Use/Account Access: identity theft or the fraudulent use of confidential personal information or account access in order to steal money, establish credit, or access account information, either through electronic or other means.

Industrial Controls and Operations: losses involving disruption or attempted disruption to "connected" physical assets such as factories, automobiles, power plants, electrical grids, and similar (including "the internet of things").

Network/Website Disruption: unauthorized use of or access to a computer or network, or interference with the operation of same, including virus, worm, malware, digital denial of service (DDOS), system intrusions, and similar.

Phishing, Spoofing, Social Engineering: attempts to get individuals to voluntarily provide information which could then be used illicitly, e.g. phishing or spoofing a legitimate website with a close replica to obtain account information, or sending fraudulent emails to initiate unauthorized activities (aka "spear phishing").

Skimming, Physical Tampering: use of physical devices to illegally capture electronic information such as bank account or credit card numbers for individual transactions, or installing

software on such point - of - sale devices to accomplish the same goal.

IT – Configuration/Implementation Errors: losses resulting from errors or mistakes which are made in maintaining, upgrading, replacing, or operating the hardware and software IT infrastructure of an organization, typically resulting in system, network, or web outages or disruptions.

IT – Processing Errors: Losses resulting from internal errors in electronically processing orders, purchases, registrations, and similar, usually due to a security or authorization inadequacy, software bug, hardware malfunction, or user error.

Cyber Extorsion: Threats to lock access to devices or files, fraudulently transfer funds, destroy data, interfere with the operation of a system/network/site, or disclose confidential digital information such as identities of customers/employees, unless payments are made.

Cybersecurity Threats, Vulnerabilities, Mitigation Measures in Industrial Control and Automation Systems: A Technical Review

Alfred Ocaka
Dept. of Electronic Engineering and Communications
Institute of Technology Carlow
Ireland
alfred.ocaka@itcarlow.ie

Diarmuid Ó Briain
Dept. of Electronic Engineering and Communications
Institute of Technology Carlow
Ireland
diarmuid.obriain@itcarlow.ie

Steven Davy
Walton Institute,
Waterford Institute of Technology
Ireland
steven.davy@waltoninstitute.ie

Keara Barrett
Dept. of Computing
Institute of Technology Carlow
Ireland
keara.barrett@itcarlow.ie

***Abstract*—Cyberattacks on Industrial Control and Automation Systems (ICAS) have significantly increased in recent years due to IT and OT convergence. Traditionally, ICAS were isolated systems running proprietary protocols on specialised software and hardware. However, to improve business processes and efficiency, ICAS vendors are adopting smart technologies such as Industrial Internet of Things (IIOT), Machine to Machine (M2M), Digital Twin, cloud computing, and Artificial Intelligence (AI). This integration presents new vulnerabilities in ICAS that can be exploited by threat actors. ICAS are utilised in critical infrastructure and widely used in power, nuclear plant, water, oil, natural gas, and manufacturing industries. Therefore, cyberattacks on these systems can pose a significant threat to humans and the environment, disrupt social services, cause financial losses, and threaten national security. Because of these threats, numerous mitigation measures are being implemented to protect ICAS from cyberattacks. However, security experience and expertise have demonstrated that we can never fully protect a system and one should never propose that their solution will fully protect. Rather one can claim that their solution / mitigation technique adds a layer to the defence in depth approach. This paper discusses the different cybersecurity standards and frameworks for ICAS, investigates the existing threats and vulnerabilities, and methods of securing ICAS**

***Index Terms*—Industrial Control and Automation Systems, Threats and Vulnerabilities in Industrial Control and Automation Systems, Industrial Cybersecurity standards and frameworks, Industrial Protocols**

I. Introduction

Industrial Control and Automation System (ICAS) is a general term that includes Process Control Systems (PCS), Distributed Control Systems (DCS), Supervisory Control And Data Acquisition (SCADA) systems, Safety Instrumented Systems (SIS), and Programmable Logic Controller (PLC). ICAS is found in critical infrastructure and is widely used in power, nuclear plant, water, oil, natural gas, and manufacturing industries [1]. The interconnections and interdependencies of the main components of ICAS are described by the Purdue Reference Model in Table 1. The hierarchical industry-adopted model has five levels, and each level presents unique cybersecurity challenges from on-site physical devices to remote connections. These challenges require defence-in-depth approach such as network segmentation, network monitoring using intrusion detection systems and physical security [2].

Level 5 - Enterprise network
Level 4: Site business planning and logistics
Industrial Demilitarized Zone
Level 3 - Site operations
Level 2 - Area supervisory control Level
Level 1 - Basic control
Level 0 - Process

TABLE I: Purdue Model

A. Level 5 - Enterprise Network

Level 5 is the enterprise network where corporate IT infrastructure systems and applications reside. The functions at this level includes enterprise resource management, Business-to-Business (B2B), and Business-to-Customer (B2C) services functions [2]. Protecting enterprise networks is challenging due to network and software vulnerabilities, and increased sophistication of attacks. The vulnerabilities and attacks at this level include network leaks, unauthorised access, backdoors, Denial of Service (DoS), direct-access attacks, eavesdropping, code injections, rootkit behaviour, and phishing [3].

B. Level 4 -Site Business Planning and Logistics

Level 4 is for Site Business Planning and Logistics which include systems that require standard access to the enterprise

network. The systems at this level include database servers, non-critical plant systems ,and other enterprise applications such as SAP. Adversary can use this level to pivot to the manufacturing zone [2]. NIST SP 800-82 [4] recommends implementing an Industrial Demilitarized Zone(IDMZ) to provide security between Level 4 and lower levels.

C. Industrial Demilitarised Zone (IDMZ)

IDMZ is a buffer that enforces data security policies between a trusted network, the Industrial Security Zones (ISZ) and an untrusted network, the Enterprise Security Zone (ESZ). It is an additional layer of defence-in-depth that provides secure data transfer between the ISZ and ESZ [2].

D. Level 3 -Site operation and management

Level 3 provides high-level monitoring and management of the industrial operations. Applications and systems at this level include batch management software, manufacturing execution/operations management systems (MES/MOMS), maintenance and plant performance management systems, Open Platform Communication (OPC) servers, and data historians. Level 3 systems and applications use standard operating systems, such as GNU/Linux or Microsoft Windows, and IT communication protocols (Transmission Control Protocol/Internet Protocol (TCP/IP), User Datagram Protocol (UDP), and Ethernet) which makes them vulnerable to IT-related attacks [2].

E. Level 2 – Supervisory control

Level 2 is comprised of devices and applications for supervising, monitoring, and controlling the physical processes. The devices at this level include standalone Human Machine Interface (HMI), DCS, and SCADA [2]. The HMI remains the main target for the adversaries since it is used by operators to monitor and control the industrial process. Compromising the HMI could facilitate attackers to manipulate the physical processes, disable alarms and notifications intended to alert operators which can be detrimental. HMI vulnerabilities include lack of authentication and authorisation, memory corruption, poor credential management, and code injection bugs. These vulnerabilities can be minimised by adhering to secure development practice [5].

F. Level 1 – Basic control

Level 1 is composed of Intelligent Devices such as PLC, Remote Terminal Unit (RTU), and Programmable Automation Controller (PAC) which control the input and output devices at Level 0 [2]. Level 1 devices are designed to ensure High Availability (HA) and efficiency but have limited security features such as authentication, encryption, and authorisation as evidenced by previous attacks such as Black Energy3 [6]. These attacks could permit an adversary to control and manipulate the input/output field devices causing physical damage to the entire industrial plant. An attacker could, for instance, replace the PLC program with a logic boom or modify the firmware [7]

G. Level 0 – Processes

This level consists of input and output devices used to measure the physical processes such as temperature, humidity, and pressure. The input and output devices include sensors, switches, valves, pumps, and motors [2].

Cyberattacks on ICAS related networks have significantly increased, in recent years. A report by Otorio [8] showed that cyberattacks on Industrial Control Systems (ICS) have significantly increased by 200% in First Quarter (Q1) of 2021 as compared to Last Quarter (Q4) of 2021. The attacks were detrimental to many manufacturing plants and critical infrastructures such as Colonial Pipeline [9], JBS Food processing company [10], and the German chemical distribution company, Brenntag [11].

This paper is divided into five sections: Section 2 reviews the common industrial cybersecurity standards and frameworks. Section 3 discusses the current threats and vulnerabilities in ICAS. Section 4 examines and analyses the various methods of securing ICAS. Finally, Section 5 presents conclusions derived from the paper.

II. Cybersecurity standards and frameworks for ICAS

Cybersecurity standards and frameworks for ICAS provide guidelines, approaches, and best practices to protect systems and networks from cyberattacks. According to Micheal [12], 47.8% of organisations in critical infrastructure sectors map their control systems to the US National Institute of Standards and Technology (NIST) Cybersecurity Framework. In addition, some of widely used frameworks include the International Society of Automation (ISA) / International Electrotechnical Commission (IEC) 62443 (32%), NIST 800-53 (31.5%), NIST 800-82 (29.6%), and ISO 27000 Series (29.1%). In recent years MITRE Adversarial Tactics, Techniques, and Common Knowledge (ATT&CK)® ICS has become increasingly popular in response to cyberattacks, especially in the oil and energy sector [12]. The majority of ICAS vendors use standards and frameworks they deem appropriate for their organisations. However, some standards are mandatory based on the sector, jurisdiction and the geographical location. For instance, North American Electric Reliability Corporation (NERC) Critical Infrastructure Protection (CIP) is mandatory for most Bulk Electrical System (BEM) vendors in the US and some other countries in North America.

A. ISA/IEC 62443

The ISA/IEC 62443 standards provide a comprehensive approach to protect IACS from cyberattacks. The holistic approach of the standards relies on the structural aspects of security strategy which include technology, process, and people.

The standard defines the roles of all the stakeholders involved in designing, developing, deploying, operating, and maintenance of IACS components and systems [13]. The product supplier is responsible for the development and commercialisation of secured components and systems. While

the system integrator is responsible for designing, deploying, and commissioning the automation solutions. Lastly, the asset owner is responsible for the specification of the requirements, operation, and maintenance of the automation solution, as well as the decommissioning of the assets at the end of their life [13].

B. NIST SP 800-82 Guide to ICS Security

The NIST Special Publication (SP) 800-82r2 describes series of recommendations, approaches, and methodologies to evaluate the security of ICS. Although NIST does not provide complete standards, it is an excellent tool to improve cybersecurity risk management in ICAS. The standard guides asset owners to identify, protect, detect, respond, and recover from cyberattacks [4] [14].

The standards, guidelines, and practices in the NIST framework provides a common taxonomy and mechanism for organisations to [15]:

- Describe their current cybersecurity posture,
- Describe their target state for cybersecurity,
- Identify and prioritise opportunities for improvement within the context of a continuous and repeatable process,
- Assess progress toward the target state,
- Communicate among internal and external stakeholders about cybersecurity risk.

C. NERC CIP Standards

NERC CIP provides a set of standards to protect BES in the North America from cyber threats using results-based approach. The result-based approach focuses on performance, risk management, and entity capabilities. Some of NERC CIP version 6 policy guidelines subject to enforcement include BES Cyber System Categorisation,Security Management Controls and Electronic Security Perimeter [16].

D. CISA Recommended Practices

The Cybersecurity and Infrastructure Security Agency (CISA) is a US agency that supports industrial vendors to comprehend and prepare for emerging ICS cyber security issues, vulnerabilities, and mitigation strategies. CISA has developed several recommendations based on cyber threats, vulnerabilities, and secure architecture design to minimise ICS cyberattacks. These recommendations cover the following areas [17];

- Improving ICS cybersecurity with Defence-in-Depth strategies,
- Creating cyber forensics plans for control systems,
- Developing an ICS Cybersecurity Incident Response Plan (IRP),
- Good practice guide for firewall deployment on SCADA and process control networks,
- Patch Management for ICS,
- Remote Access for ICS,
- Updating Anti-virus in an ICS,
- Mitigations for vulnerabilities in ICS networks.

E. MITRE ATT&CK for Industrial Control Systems

MITRE ATT&CK for ICS (MAICS) provides various phases of an adversary's attack life cycle, as well as the assets and systems they are known to target. These tactics include initial access, execution, persistence, evasion, discovery, lateral movement, collection, command, and control, inhibit response function, impair process control, and impact. MAICS techniques can apply to common ICS components which include control server, data historian, engineering workstation, field controller/RTU/PLC/IED, HMI and Safety Instrumented System (SIS). Majority of these devices are in levels 1 and 2 of the Purdue Model [18].

The MAICS's initial access techniques include drive-by compromise, spearphishing, supply chain compromise, and replication through removable media. After gaining access to ICS network, a threat actor can use techniques such as hooking or command-line interface to execute malicious code. In the persistence phase, a threat actor can use valid account to perform privilege escalation and modification of system's application or firmware [18].

III. Vulnerabilities and Threats in ICAS

A. Threats in ICAS

The threat landscape in ICAS has drastically changed in recent years. A report by [19] showed that "46% of known OT threats are poorly detected or not detected at all". The poor detection rate could be attributed to several reasons such as increase in the sophistication of attacks with capability of evasive behaviours. Threats to ICAS can be classified into adversarial, environmental (natural and man-made disasters), accidental and structural [4]. By understanding threat sources and attributes such as capability and intention, ICAS vendors can minimise attacks. Table II summarises major attacks on ICAS systems from 2016 to 2021. The most imminent threats to ICAS are adversarial and environmental.

Adversaries such as nation states, hacktivist, organised criminal groups, script kiddies and hackers pose a significant threat to ICAS. There have been several previous attacks on ICAS by adversaries such as attacks on Colonial Pipeline [20], JBS Company [10] and the Oldsmar water treatment system [21]. In the Oldsmar's attack, the adversary attempted to poison the water system by changing the level of sodium hydroxide to more than 100 times its normal level. Fortunately, the operator detected the changes and immediately corrected the level of sodium hydroxide. If the attack was not detected early, it would have caused health related complications to a city of about 15,000 people [21].

Natural and man-made disasters such as hurricanes, earthquakes, bombing and flood/tsunami pose an immense threat to ICAS and critical infrastructure. In 2020 alone, a total of 313 major natural disasters occurred worldwide [22]. The magnitude of these disasters on ICAS could not fully be quantified, thus not included in Table 2. In order to reduce the impact of natural disasters, it is essential to implement robust and resilient systems [23].

Furthermore, ICAS are affected by unintentional threats which are mostly accidental, such as incorrect configuration of the industrial systems during installation, maintenance, and other human error during daily operations. Previous incidents caused by human error and breaches of safety principles include the Tokaimura Nuclear accident (1999) in Japan and the Chernobyl accident (1986) in Ukraine. Adequate training of employees and designing a comprehensive safety polices can minimised these threats [24].

B. Vulnerabilities in ICAS

ICAS has weaknesses and flaws that can be exploited by threat actors. These vulnerabilities exist in the hardware, firmware, software application, network, and the process [7]. According to Claroty [29] report, the number of ICAS vulnerabilities published significantly increased from 449 in 2nd of 2020 to 637 in the 1st Quarter of 2021. The largest percentage of vulnerabilities disclosed affected Level 3 of the Purdue Model: Operations Management (23.55%), followed by Level 1: Basic Control (15.23%) and Level 2: Supervisory Control (14.76%). Remotely exploitable vulnerabilities attributed to 61.38% lower than 71.49% reported in the 2nd Quarter of 2020. Furthermore, vulnerabilities exploitable through local attack vectors in the 1st Quarter of 2021 rose to 31.55% from 18.93% in the 2nd Quarter 2020. The vulnerabilities discovered by sources external to the affected vendor, including research organisations, independent researchers, and academics, among others attributed to 80.85%.

1) Hardware: ICAS hardware components such as PLCs, RTUs, HMI, and PAC are vulnerable to cyberattacks. According to Basnight et al. [30], the three main hardware attack vectors include physical manipulation of the hardware, software exploitation of the hardware designed flaws, and supply chain compromise. The physical manipulation requires physical access to the hardware, for instance through or by malicious insiders. In addition, supply chain compromise involves an attacker infiltrating the manufacturing process to create vulnerabilities or backdoors [30].

2) Firmware: The firmware modification attack of field devices such as PAC, PLC, and RTU requires vulnerable analysis which is majorly implemented through reverse engineering [30]. Several researchers have demonstrated how firmware vulnerabilities can be exploited by adversaries. Basnight et al. [30] demonstrated how legitimate firmware can be updated and uploaded to an Allen-Bradley ControlLogix L61 PLC. Similarly, Konstantinou and Maniatakos [31], presented the impact of firmware modification attacks on power systems field devices. They demonstrated how to reverse engineer the firmware of a relay controller and inject malicious tripping commands to disrupt the operation of Circuit Breakers. Hui, McLaughlin and Sezer [32], used reverse engineering of the software and communication protocols used by Siemens S7-1211C PLCs. They demonstrated how an attacker can take control of the PLC using the potential exploits of the communication protocol.

3) Software: ICAS systems such as HMI, engineering workstations, historians, and Open Platform Communications (OPC) servers use software and applications which are susceptible to cyberattacks. The software stack are vulnerable to buffer overflows, SQL injection, and cross-site scripting due to poor input validation, violation of least privilege, and other access control errors [33]. In addition, some of these vulnerabilities are zero-days without any security patches. The famous attack that used software vulnerabilities to access the industrial network was Stuxnet. The malware exploited four zero-day vulnerabilities that existed in Microsoft Windows operating system, at the time, to gain access to Purdue level 0, 1 and 2 devices at the Iranian nuclear facility at Natanz [34].

4) Network: Network vulnerabilities are associated with several factors such as insecure network architecture design, unencrypted communication protocols, insecure device configuration and lack of network security management policy [35]. For instance, implementation of a flat ICAS network with no zones, no port security, and lax enforcement of remote access policies can lead to the entire network being compromised [36]. Cybersecurity standards and frameworks such as ISA/IEC 62443 and NIST 800-82r2 provide guidelines and best practices to secure industrial networks.

5) Industrial Communication Protocols: Industrial communication protocols (ICP) connect various systems and instruments in the ICAS domain. ICP are designed to be efficient and reliable to meet the real-time communication requirement of the industrial network [1]. Therefore, majority of ICPs are insecure by design to conform to the industrial environment. These vulnerabilities include lack of authentication, authorisation, and encryption. Thus, ICPs are vulnerable to attacks such as Man-in-the-Middle (MitM), DoS, replay, injection, spoofing, and eavesdropping [37].

Modbus: Modicon Communication Bus (Modbus) is an industrial protocol that operates in a master-slave or server-client model. The master or server initiates the request, and the slave or client responds to it. The standard Modbus protocol lacks basic security features such as authentication and encryption to ensure high efficiency, stability, and reliability of the industrial environment. In addition, the protocol does not have any inherent checksum or integrity checking mechanisms, making it susceptible to flooding, spoofing, and replay attacks [1] [38].

DNP3: Distributed Network Protocol 3 (DNP3) is a non-proprietary communication used in SCADA and remote monitoring systems. The protocol facilitates communication between the master station and the substations. It is widely used in the water, oil, and gas sectors. The protocol was designed to ensure high reliability and efficiency but lacks security mechanisms such as authentication and encryption [39] [40].

Thus, the protocol is vulnerable to attacks such as MitM, packet modification, and injection, DoS attack, replay attack and spoofing. DNP3 Secure Authentication (DNP3-SA) has been developed to address the security weakness of DNP3. It introduces a separate protocol layer between the DNP3 Application Layer and the DNP3 transport layer which can address

Incident	Year	Threat Actor	Affected Areas	Initial Access point	Sector	Impact
Ukrainian Power Grid (Industroyer) [25]	2016	Adversarial (Organised group)	Ukraine	Spear phishing	Energy	Disk wipe, loss of productivity and revenue, loss of safety
Wolf Creek Nuclear Operating Corporation [25]	2017	Adversarial (Organised group)	Kansas, USA	Spearphishing	Civil nuclear	Not disclosed
Cadbury Factory Attack [25]	2017	Adversarial (Organised group)	Australia	External remote service	Food	Food Loss of productivity and revenue
Triton [25]	2017	Adversarial (Nation state)	Saudi Arabia	Workstation compromise	Petrochemical	Denial of control, Loss of safety
Norsk Hydro [25]	2019	unknown	Norway	Spear phishing	Manufacturing and energy	Loss of view
Kansas Water Treatment Plant [8]	2019	Adversarial (Insider)	USA	Remote Access services	Water	Not disclosed
Shahid Rajaie Port Attack [25]	2020	Adversarial (Nation state)	Iran	Unknown	Transport	Loss of productivity and revenue
Honda Factories Attack [25]	2020	unknown	USA, Turkey	Spear phishing	Manufacturing	Denial of control
CPC Corporation [26]	2020	unknown	Taiwan	unknown	Petrochemical	Loss of revenue
Oldsmar water treatment system [21]	2021	Adversarial	Oldsmar, Florida, US	Remote Access (Team Viewer)	Water	Increased level of sodium hydroxide by 100 times
Colonial Pipeline Company [27]	2021	Adversarial (Organised criminal group)	USA	VPN	Petroleum and gas	Loss of revenue and productivity
JBS Company [28]	2021	Adversarial (Organised criminal group)	USA, Australia; Canada; Brazil	Not disclosed	Food	Loss of revenue and productivity

TABLE II: Summary of major attacks

four threats: spoofing, modification, replay, and eavesdropping [40].

ICCP/IEC 60870-6: Inter-Control Centre Communications Protocol (ICCP) was designed to provide data exchange over Wide Area Network (WAN) between utility control centres, Independent System Operators (ISO), Regional Transmission Operators (RTO), and other Generators [41]. ICCP is vulnerable to session hijacking and spoofing attacks because it lacks authentication and encryption security features. Similarly, the protocol is susceptible to attacks such as MITM, DoS, and Distributed DoS (DDoS) since it provides data exchange over WAN [42]. The inherent vulnerability of standard ICCP led to the development of Secure ICCP that uses Transport Layer Security (TLS) to provide encryption and authentication [41].

OPC: Open Platform Communication (OPC) industrial protocol is primarily designed to provide communication between Personal Computer (PC) based software and automation devices. Its design was based on Object Linking and Embedding (OLE) which works on client/server mode [43]. OPC is susceptible to different forms of attack such as Buffer Overflow (BOF) and DoS due to the use of Distributed Component Object Model (DCOM) and Remote Procedure Call (RPC) [44]. OPC United Architecture (OPC-UA) addresses the security concerns of OPC and provides greater interoperability, eliminating the Microsoft Windows dependency, but maintaining retro compatibility with its predecessor [45].

IV. Securing ICAS

Securing ICAS from cyberattacks requires implementation of layers of security measures. There is no single solution that can prevent all ICAS attack vectors. Therefore, ICAS vendors should implement comprehensive security measures to protect the physical assets, networks, devices, and software applications [2].

A. Intrusion Detection System

Intrusion Detection System (IDS) monitors ICAS networks, systems and detect any malicious activity or abnormal behaviour. IDS detects malicious activity by collecting and analysing various data sources such as network traffic, security logs, audit data, and system/application. Furthermore, IDS can detect network related attacks such as DoS, MITM and other forms of malware in ICAS environment [46].

Conventional IDS are classified into Misuse-based (MIDS) and Anomaly-based IDS (AIDS). MIDS uses known signatures to detect attacks, while AIDS uses statistical and Machine Learning (ML) algorithms to detect anomalies. In addition, IDS can be categorised based on the data sources; Host-based IDS (HIDS) and Network-based (NIDS). HIDS monitors a specific host for malicious activities and very effective to detect advanced persistent threats on a specific host [47]. NIDS detects anomalies in the entire network; however, it lacks visibility into the internal nodes and cannot locate the specific node under attack [48] [4]. In ICAS, IT systems at the Purdue Levels 2 and 3 can be monitored using open-source IDS such as Suricata, Snort and Zeek. While industrial zone requires OT-oriented IDS (OT-IDS) such as Forescout's eyeInsight [49].

1) Misuse or signature-based IDS: Misuse IDS identifies abnormal behaviour of a system or network based on the signatures of the known attacks and vulnerabilities. MIDS has a very high detection rate and effective in detecting known

attacks. However, MIDS cannot detect unknown attacks, such as zero-days, whose signatures have not yet been included in the database. This is challenging with the increasing zero-day vulnerabilities being discovered in ICAS. A good example of an open source MISD for monitoring Purdue model level 3 devices is Snort [50] [49].

2) Anomaly-based IDS: AIDS uses a defined normal behaviour baseline to detect any malicious activity in a system or a network. The IDS raises an alert when the deviation between the current behaviour and the normal behaviour transcends the predefined threshold. AIDS can identify a variety of unknown attacks but has a relatively high false alarm rate [47]. AIDS can mainly be classified into the statistical base and machine learning-based approaches. Statistical-based IDS uses statistical methods and algorithms such as Time series analysis, and Markov chain to process system events and network traffic [51]. While ML-based IDS uses ML algorithms to classify the normal and abnormal behaviour of the system or network. Some of the ML algorithms include Support Vector Machines (SVM), deep learning, clustering and classification, and decision trees [46].

Industrial network behaviour tends to be highly predictable especially in well-isolated control zones, making anomaly detection more reliable [1]. There are several off-the-self AIDS for OT systems such as Darktrace for OT, Dragos Platform, Forescout and Claroty Platform [52].

B. Network segmentation

Network segmentation involves the logical grouping of information systems and ICAS devices. It provides an additional layer of defence against cyberattacks [2]. NIST recommends separating the ISZ from the ESZ by creating an Industrial Demilitarised zone (IDZ) [4]. In addition, ICAS networks can be further segmented into small sub-networks called zones. Each zone is protected by security perimeter that monitors and filters network traffic. Although network segmentation is very crucial for securing industrial networks, it can be challenging to implement for complex networks with many interconnected devices [1].

The network segmentation approach includes physical implementation, logical implementation, and network traffic filtering [53]. Physical segmentation involves partitioning the ICAS network infrastructure into small physical components. It guarantees high security but costly since it involves investing in additional hardware such as routers, gateways and switches [54]. Logical segmentation can be implemented using Virtual Local Area Networks (VLAN) or Virtual Private Network (VPN). It is cost-effective but relatively insecure compared to physical segmentation. Several previous ICAS network-related attacks exploited VPN vulnerabilities. For instance, in the Colonial Pipeline attack, the adversaries used a legacy VPN to gained access to the IT system of the company [27]. Lastly, network traffic filtering provides segmentation by restricting certain parts of the system from communicating with others [53].

C. Securing endpoints

Securing devices in ICAS requires holistic security approaches such as patch management, anti-malware software, device hardening, application whitelisting, monitoring and logging, least user privilege, password management, configuration, and change management [2] [55].

1) Patch management : Patch management involves identification, installation, scheduling, and verification of the patches to be applied to a system to either add new functionalities or resolve vulnerabilities [56]. ICAS has HA requirements, and any unplanned downtime can cause physical damage, financial losses and other safety concerns. Therefore, a systematic patch management process such as rigorous testing of the patches is required before deployment. This is to ensure safety and minimise risk of downtime. In addition, patching ICAS legacy systems can be challenging since most vendors no longer support these systems [4].

2) Antimalware software: Anti-malware such as antivirus is relatively effective in defending against malware in both ESZ and the ISZ when properly installed, configured, and maintained. Adoption of antivirus in ICAS requires special precautions such as compatibility checks, change management issues, and performance impact metrics. In addition, regressive testing of the antimalware is required by both the vendor and the asset owner to avoid disruption of the industrial operation [4] [57].

3) Application WhiteListing: Application WhiteListing (AWL) involves setting a list of approved applications that are allowed while blocking any application not on the list [58]. AWL can be applied to detect and prevent attempts to execute malware uploaded by adversaries on ICAS devices such as HMIs, PLCs, RTUs and servers. The downside of AWL implementation is its inability to effectively detect malware that exploit applications that run in the higher-level execution environments, such as Java, .NET Framework, and other scripting languages. In addition, AWL is pointless in an event where the adversary has already gained access and escalated the privileges in the ICAS environment. The adversary can modify the AWL list to allow any executable malware without being flagged [59].

4) Device Hardening : The suppliers and manufacturers play a critical role in securing ICAS devices. Some of the devices such as PLCs have hard-coded credentials which can easily be extracted by reverse engineering the firmware [2]. For instance, the recent hard-coded key vulnerability (CVE-2021-22681, CVSS 10.0) which affected Rockwell Automation Logix PLCs [60].

5) Configuration/change management: Integrating ICAS cyber security and reliability requires proper configuration of all assets, including operating systems, networking equipment, and other embedded devices. Improper configuration practices such as the use of device default passwords, and plain text transmission of usernames and passwords can allow adversaries to gain full access to the industrial network [35]. The NERC CIP-010-4 provides configuration change management

and vulnerability to prevent and detect unauthorised changes to BES [61].

D. Physical security

Physical security controls involve the implementation of physical measures to limit an authorise access to the ICAS environment. Unauthorised physical access to the ICAS environment can lead to physical modification, theft, and destruction of industrial systems which can have a devastating impact [62]. NERC CIP and ISA/IEC 62443 provide recommendations, and guidelines to protect the ICAS physical assets. Some of these physical protection settings include physical boundaries such as a fence, a closed control house, locked cabinets, and installing video cameras for monitoring purposes.

E. Cybersecurity awareness and training

Cybersecurity awareness and training support employees to recognise potential threats and take the appropriate actions to minimise cyber risks. Employees should have basic knowledge on cybersecurity such as phishing, physical access, password management, and organisation cybersecurity policies and guidelines [63]. Several previous ICAS attacks such as the Honda Factories attack, the Black Energy3 and the Wolf Creek Nuclear Operating Corporation attack used spearphishing as initial access vector [25].

F. Risk assessment

A comprehensive risk assessment in ICAS helps to identify, categorise, prioritise, and mitigate risk. Risk assessment involves asset identification, system characterisation, vulnerability identification, threat modelling, risk calculation, and mitigation planning [2] [13].

V. Conclusion

Cyberattacks on ICAS are on the rise due to several factors such as OT/IT convergence and threats from nation-states, organised criminal groups, and hacktivists. This paper provided an overview of the current unique cybersecurity challenges at each level of the Purdue Reference Module. As noted, securing devices at Purdue level 0 and 1 remains a challenge, since majority of these devices have limited resources and lack strong security measures such as authentication. Furthermore, this paper examined the widely used security standards and frameworks in ICAS. Though most of these standards provide similar guidelines, the NIST framework and ISA/IEC 62443 are widely adopted by critical infrastructure organisations. The current threats and vulnerabilities in the ICAS were also considered. Adversarial and environmental threats pose high risk to ICAS. Lastly, this paper recommends approaches and best practices to secure ICAS. Further research and investigations are still required to protect ICAS from the evolving cyberattacks. Future work in this project will focus on exploring methodologies and approaches to secure Purdue model level 0, 1, and 2 devices.

References

[1] D. K. Eric and J. T. Langill, *Industrial Network Security*, second edi ed. Elsevier, 2015.

[2] P. Ackerman, *Industrial Cybersecurity*, 2017, no. 11.

[3] E. U. Opara and O. J. Dieli, "Enterprise Cyber Security Challenges to Medium and Large Firms : An Analysis," *I.J. ofElectronics and Information Engineering*, vol. 13, no. 2, pp. 77–85, 2021.

[4] K. Stouffer, V. Pillitteri, S. Lightman, M. Abrams, and A. Hahn, *Guide to Industrial Control Systems (ICS) Security*, 2015, vol. 2. [Online]. Available: http://dx.doi.org/10.6028/NIST.SP.800-82r2

[5] B. Gorenc and F. Sands, "Hacker Machine Interface: The State of SCADA HMI Vulnerabilities," p. 30, 2017. [Online]. Available: https://documents.trendmicro.com/assets/wp/wp-hacker-machine-interface.pdf?_ga=2.96440475.667063318.1554618734-319631345.1554618734

[6] J. E. Sullivan and D. Kamensky, "How cyber-attacks in Ukraine show the vulnerability of the U.S. power grid," *Electricity Journal*, vol. 30, no. 3, pp. 30–35, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.tej.2017.02.006

[7] S. McLaughlin, C. Konstantinou, X. Wang, L. Davi, A. R. Sadeghi, M. Maniatakos, and R. Karri, "The Cybersecurity Landscape in Industrial Control Systems," *Proceedings of the IEEE*, vol. 104, no. 5, pp. 1039–1057, 2016.

[8] C. Otorio, "INDUSTRIAL CYBERCRIME IMPACT Q1 2021 REPORT," *OTORIO*, no. April, pp. 1–17, 2021. [Online]. Available: https://f.hubspotusercontent20.net/hubfs/8127371/R_Industrial-Cybercrime-Impact-Report-Q1-2021-21-04-22.pdf

[9] A. Greenberg, "The Colonial Pipeline Hack Is a New Extreme for Ransomware," 2021. [Online]. Available: https://www.wired.com/story/colonial-pipeline-ransomware-attack/

[10] B. Menegus, "World's largest meat supplier grinds to a halt after cyberattack," jun 2021. [Online]. Available: https://www.theverge.com/2021/6/1/22463621/jbs-cyberattack-russia-largest-meat-supplier

[11] L. Abrams, "Chemical distributor pays $4.4 million to DarkSide ransomware," may 2021. [Online]. Available: https://www.bleepingcomputer.com/news/security/chemical-distributor-pays-44-million-to-darkside-ransomware/

[12] D. Micheal, "Understanding OT Frameworks and Standards for a More Secure Industrial Network," oct 2021. [Online]. Available: https://www.garlandtechnology.com/blog/understanding-ot-frameworks-and-standards-for-a-more-secure-industrial-network

[13] P. Kobes, *Guideline Industrial Security : IEC 62443 is easy*, 2017.

[14] A. A. Jillepalli, F. T. Sheldon, D. C. De Leon, M. Haney, and R. K. Abercrombie, "Security management of cyber physical control systems using NIST SP 800-82r2," *2017 13th International Wireless Communications and Mobile Computing Conference, IWCMC 2017*, pp. 1864–1870, 2017.

[15] N. Barrett, Matt, "Framework for improving critical infrastructure cybersecurity," *Proceedings of the Annual ISA Analysis Division Symposium*, vol. 535, pp. 9–25, 2018.

[16] NERC, "NERC." [Online]. Available: https://www.nerc.com/Pages/default.aspx

[17] CISA, "Recommended Practices — CISA." [Online]. Available: https://us-cert.cisa.gov/ics/Recommended-Practices

[18] O. Alexander, M. Belisle, and J. Steele, "MITRE ATT&CK ® for Industrial Control Systems: Design and Philosophy," no. March, pp. 1–43, 2020.

[19] Honeywell, "INDUSTRIAL USB THREAT REPORT 2021," 2021.

[20] T. W. Mike Hoffman, "Recommendations Following the Colonial Pipeline Cyber Attack," 2021. [Online]. Available: https://www.dragos.com/blog/industry-news/recommendations-following-the-colonial-pipeline-cyber-attack/

[21] V. Amir, L. Jamiel, and C. Christina, "Florida water system hack: Someone tried to poison Oldsmar city with sodium hydroxide, sheriff says - CNN," feb 2021. [Online]. Available: https://edition.cnn.com/2021/02/08/us/oldsmar-florida-hack-water-poison/index.html

[22] UNDRR, "Global Natural Disaster Assessment Report 2020," *United Nations Office for Disaster Risk Reduction (UNDRR)*, no. October, pp. 1–45, 2021.

[23] A. Urlainis, I. M. Shohet, R. Levy, D. Ornai, and O. Vilnay, "Damage in critical infrastructures due to natural and man-made extreme events - A critical review," *Creative Construction Conference 2014*, vol. 85, pp. 529–535, 2014.

[24] L. Chiara and A. Amendola, "Human Errors Analysis and Safety Management Systems in Hazardous Activities," *... for Applied Systems Analysis ...*, 2005. [Online]. Available: http://webarchive.iiasa.ac.at/Publications/Documents/IR-05-003.pdf

[25] T. Miller, A. Staves, S. Maesschalck, M. Sturdee, and B. Green, "Looking back to look forward: Lessons learnt from cyber-attacks on Industrial Control Systems," *International Journal of Critical Infrastructure Protection*, vol. 35, no. May, p. 100464, 2021. [Online]. Available: https://doi.org/10.1016/j.ijcip.2021.100464

[26] Cyberint, "Targeted Ransomware Attacks in Taiwan - Cyberint," may 2020. [Online]. Available: https://cyberint.com/blog/research/targeted-ransomware-attacks-in-taiwan/

[27] K. Stephanie and R.-a. Jessica, "One password allowed hackers to disrupt Colonial Pipeline, CEO tells senators," 2021. [Online]. Available: https://www.reuters.com/business/colonial-pipeline-ceo-tells-senate-cyber-defenses-were-compromised-ahead-hack-2021-06-08/

[28] M. Bryan, "World's largest meat supplier grinds to a halt after cyberattack," 2021. [Online]. Available: https://www.theverge.com/2021/6/1/22463621/jbs-cyberattack-russia-largest-meat-supplier

[29] Claroty Team82, "INDUSTRIAL CYBERCRIME IMPACT Q1 2021 REPORT," *Claroty*, pp. 1–45, 2021. [Online]. Available: https://security.claroty.com/biannual-ics-risk-vulnerability-report-2H-2020

[30] Z. H. Basnight, "Firmware Counterfeiting and Modification Attacks on Programmable Logic Controllers," p. 120, 2013.

[31] C. Konstantinou and M. Maniatakos, "Impact of firmware modification attacks on power systems field devices," *2015 IEEE International Conference on Smart Grid Communications, SmartGridComm 2015*, pp. 283–288, 2016.

[32] H. Hui, K. McLaughlin, and S. Sezer, "Vulnerability analysis of S7 PLCs: Manipulating the security mechanism," *International Journal of Critical Infrastructure Protection*, vol. 35, no. September 2019, p. 100470, 2021. [Online]. Available: https://doi.org/10.1016/j.ijcip.2021.100470

[33] Q.-q. WU, L.-h. WEI, Z.-q. LIANG, Z.-w. YU, M. CHEN, Z.-h. CHEN, and J.-j. TAN, "Patching Power System Software Vulnerability Using CNNVD," *2018 International Conference on Computer, Communications and Mechatronics Engineering (CCME 2018)*, pp. 356–360, 2018.

[34] S. Kriaa, M. Bouissou, and L. Piètre-Cambacédès, "Modeling the Stuxnet attack with BDMP: Towards more formal risk assessments," *7th International Conference on Risks and Security of Internet and Systems, CRiSIS 2012*, pp. 1–8, 2012.

[35] W. Richard, "How to Design and Configure Secure Industrial Networks," 2017. [Online]. Available: https://blog.isa.org/how-to-design-configure-secure-industrial-networks

[36] Rockwell & CISCO, "Securely Traversing IACS Data across the Industrial Demilitarized Zone," no. May, 2017.

[37] X. Yikai, Y. Yi, L. Tianran, J. Jiaqi, and W. Qi, "Review on Cyber Vulnerabilities of Communication Protocols in Industrial Control Systems," pp. 1–6, 2017.

[38] P. Huitsing, R. Chandia, M. Papa, and S. Shenoi, "Attack taxonomies for the Modbus protocols," *International Journal of Critical Infrastructure Protection*, vol. 1, no. C, pp. 37–44, 2008. [Online]. Available: http://dx.doi.org/10.1016/j.ijcip.2008.08.003

[39] M. Majdalawieh, F. Parisi-Presicce, and D. Wijesekera, "DNPSec: Distributed network protocol version 3 (DNP3) security framework," *Advances in Computer, Information, and Systems Sciences, and Engineering - Proceedings of IETA 2005, TeNe 2005, EIAE 2005*, vol. 3, pp. 227–234, 2006.

[40] B. Sangewar and A. R. Buchade, "Survey On Analysis Of Security Threats In DNP3 Protocol," vol. 9, no. 06, pp. 365–369, 2020.

[41] J. T. Michalski, A. Lanzone, J. Trent, and S. Smith, "Secure ICCP Integration Considerations and Recommendations," no. June, pp. 1–98, 2007.

[42] M. Nitesh, "TASE 2.0 and ICCP - Infosec Resources," feb 2020. [Online]. Available: https://resources.infosecinstitute.com/topic/tase-2-0-and-iccp/

[43] Q. Wanying, W. Weimin, Z. Surong, and Z. Yan, "The Study of Security Issues for the Industrial Control Systems Communication Protocols," no. Jimet, pp. 693–698, 2015.

[44] B. Rolston, "Security Implications of OPC , OLE , DCOM , and RPC in Control Systems," no. January, p. 254795, 2006. [Online]. Available: https://inldigitallibrary.inl.gov/sites/sti/sti/3494180.pdf

[45] M. Conti, D. Donadel, and F. Turrin, "A Survey on Industrial Control System Testbeds and Datasets for Security Research," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2021.

[46] M. Kaouk, J. M. Flaus, M. L. Potet, and R. Groz, "A review of intrusion detection systems for industrial control systems," *2019 6th International Conference on Control, Decision and Information Technologies, CoDIT 2019*, pp. 1699–1704, 2019.

[47] Y. Hu, A. Yang, H. Li, Y. Sun, and L. Sun, "A survey of intrusion detection on industrial control systems," *International Journal of Distributed Sensor Networks*, vol. 14, no. 8, 2018.

[48] Checkpoint, "What is an Intrusion Detection System (IDS)? - Check Point Software." [Online]. Available: https://www.checkpoint.com/cyber-hub/network-security/what-is-an-intrusion-detection-system-ids/#

[49] P. Ackerman, *Industrial Cybersecurity Second Edition*, second edi ed. Packt Publishing, 2021.

[50] R. Samrin and D. Vasumathi, "Review on anomaly based network intrusion detection system," *International Conference on Electrical, Electronics, Communication Computer Technologies and Optimization Techniques, ICEECCOT 2017*, pp. 141–147, 2017.

[51] S. Jose, D. Malathi, B. Reddy, and D. Jayaseeli, "A Survey on Anomaly Based Host Intrusion Detection System," *Journal of Physics: Conference Series*, vol. 1000, no. 1, 2018.

[52] C. M. Hurd and M. V. McCarty, *A Survey of Security Tools for the Industrial Control System Environment*, 2017, no. 571. [Online]. Available: https://www.osti.gov/servlets/purl/1376870

[53] R. Arief, N. Khakzad, and W. Pieters, "Mitigating cyberattack related domino effects in process plants via ICS segmentation," *Journal of Information Security and Applications*, vol. 51, p. 102450, 2020. [Online]. Available: https://doi.org/10.1016/j.jisa.2020.102450

[54] P. Steve, "Network Segmentation: What it is & How it Works," 2019. [Online]. Available: https://www.auvik.com/franklyit/blog/network-segmentation/

[55] C. Davidson, T. Andel, M. Yampolskiy, T. McDonald, B. Glisson, and T. Thomas, "On SCADA PLC and fieldbus cyber-security," *Proceedings of the 13th International Conference on Cyber Warfare and Security, ICCWS 2018*, vol. 2018-March, no. March, pp. 140–148, 2018.

[56] U. Gentile and L. Serio, "Survey on international standards and best practices for patch management of complex industrial control systems: The critical infrastructure of particle accelerators case study," *International Journal of Critical Computer-Based Systems*, vol. 9, no. 1-2, pp. 115–132, 2019.

[57] V. Matvey, "Updating antimalware solutions in industrial control systems," 2018. [Online]. Available: https://www.kaspersky.com/blog/updating-in-ics/20962/

[58] T. William, *Cybersecurity for SCADA Systems*, 2nd ed. Pennwell Books, 2020.

[59] DHS and NCCIC, "Guidelines for Application Whitelisting in Industrial Control Systems," *Cybersecurity and Infrastructure Security Agency (CISA)*, pp. 1–6. [Online]. Available: https://us-cert.cisa.gov/sites/default/files/documents/Guidelines for Application Whitelisting in Industrial Control Systems_S508C.pdf

[60] B. Sharon, "CLAROTY DISCOVERS CRITICAL AUTHENTICATION BYPASS IN ROCKWELL SOFTWARE," 2021. [Online]. Available: https://claroty.com/2021/02/25/blog-research-critical-authentication-bypass-in-rockwell-software/

[61] NERC, "CIP-010-4 – Cyber Security — Configuration Change Management and Vulnerability Assessments," *North American Electric Reliability Corporation(NERC)*.

[62] Nccic, Ics-cert, and DHS, "Recommended Practice: Improving Industrial Control System Cybersecurity with Defense-in-Depth Strategies Industrial Control Systems Cyber Emergency Response Team," *Communications Integration Center (NCCIC)*, 2016.

[63] A. Nagarajan, J. M. Allbeck, A. Sood, and T. L. Janssen, "Exploring game design for cybersecurity training," *Proceedings - 2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems, CYBER 2012*, pp. 256–262, 2012.

An Analysis of Ireland's Homecare Companies' Cookie Practices in terms of GDPR Compliance.

Gerard Reynolds
Department of Computing, School of Science and Computing
Galway Mayo Institute of Technology
Castlebar Co Mayo, Ireland
gerard.reynolds@aghra.ie

Seamus Dowling
Department of Computing, School of Science and Computing
Galway Mayo Institute of Tecnology
Castlebar Co Mayo, Ireland
seamus.dowling@gmit.ie

***Abstract*—The General Data Protection Regulation (GDPR) 2016, the Data Protection Act 2018 and the e-Privacy Directive 2002 are applicable legal instruments that impose responsibilities on homecare companies as 'data controllers'. One of these responsibilities is that they provide their clients with information pertaining to their clients' rights and the data controllers' responsibilities as set out by the GDPR. Many homecare companies publish their Privacy Policies on the company website to showcase their compliance and make themselves more attractive to potential clients. Many websites use Cookie (or consent) Management Platforms (CMP's) to manage their cookies and fulfil their legislative obligations. Cookies gather information and must comply with the terms of data protection and e-Privacy legislation. This research evaluates homecare companies' Cookie Practices to ascertain GDPR compliance and found them to be lacking the substance and detail necessary to be considered compliant. This was achieved by identifying the websites of homecare companies operating in Ireland, accessing their website using a cookie cleared browser and then examining the researcher's computer immediately afterwards to see what (if any) cookies had been uploaded, in addition to assessing the homecare companies CMP (where present) for compliance. This research found a high level of non-compliance and suggests that Ireland's Data Protection Commission (DPC) could and should become more involved in creating solutions by evolving their role to that of a Data Protection Service Provider. By doing so they will improve compliance with data protection legislation and enhance the protections afforded to an individual's right to privacy.**

Keywords—data protection, GDPR, cookies, privacy

I. INTRODUCTION

The Health Service Executive's National Service Plan identified that 19.2 million home support hours are to be delivered to 53,700 people.[1] Homecare companies have become increasingly reliant on information technology to manage and coordinate the delivery of these home support hours. The e-Privacy regulation (S.I.No.336/2011)[2], in addition to GDPR Recital 30 [3], aims to ensure electronic communications are conducted in such a way as to protect a person's right to privacy. [1] The e-Privacy Directive acknowledges that IP addresses and cookie identifiers might, when combined with other unique identifiers, be used to profile natural persons and by doing so breach the right to privacy. This paper and associated research focuses on GDPR compliance in the Irish homecare environment. In doing so it achieves the following:

- identified and consolidated the requirements that the GDPR and the e-Privacy Directive impose on homecare companies operating in Ireland specific to their websites' use of Cookies.
- develops a criterion matrix for evaluating homecare companies' Privacy Policies and Cookie Practices.
- provides an indication of the level of compliance amongst homecare companies in Ireland.
- poses the question; *are Homecare Companies' cookie practices GDPR compliant?* And answers it by collating and analysing homecare companies' website cookie practices.

II. RELATED RESEARCH

A. Cookies

A cookie is a small text file that may be stored on a computer or mobile device that contains data relating to a visited website. It may allow a website to 'remember' an individual's actions or preferences over a period of time, or it may contain data relating to the function or delivery of the site. It is considered valuable information as it can be used to identify an individual's interests or preferences. This information can subsequently be used by those involved in sales and marketing.

A CMP is a tool that controls user consent on websites, it requires a site user/visitor to give their consent to their data being collected managed via cookies. They should be built to comply with the latest data privacy legislation, by providing the necessary information in an appropriate manner to enable the site user/visitor to make an informed decision.

Yang et al [4] assert that data governance involves coordination of people, policies, processes, strategies, standards, and technologies to allow organisations to utilise data. CompTia [5], conducted an online survey of companies based in the United States, (425 businesses responded) and their report identified 'privacy concerns' as one of the main driving forces behind investment and expenditure in cybersecurity. Their report highlighted the importance of Governance Risk and Compliance (GRC) as functions essential to cybersecurity. Their report concluded that GRC is less technical but more reliant upon an understanding of the regulatory environment. The legislative environment is complex and evolving, the GDPR [3] contains 7 principles, 99 Articles and 173 Recitals, SI 336 /2011 [2] has 35 Sections and the Data Protection Act [6] contains 7 Parts 156 Sections and 67 amendments and case law continues to evolve. However, lack of understanding is not considered an acceptable excuse by the Courts or the DPC for

[1] Directive 2002/58/EC Of the European Parliament and of the Council 12 July 2002.

noncompliance. Whilst the regulatory environment may be considered complex,

Article 12 (1) requires that homecare companies acting as data controllers (as they are the owners of their websites) '*take appropriate measures to provide any information relating to processing to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language*'. Proof of compliance can be established using an appropriate Privacy Policy and via CMP.

Ireland's e-Privacy Regulations SI 336/2011 [2] transposed the EU e-Privacy Directive 2002/58/EC (amended in 2009) into Irish law to protect the confidentiality of electronic communications, including the use of cookies and similar technologies.

Regulation 5(3) of the e-Privacy directive is clear in stating the rules regarding use of cookies and states: A person shall not use an electronic communications network to store information, or to gain access to information already stored in the terminal equipment of a subscriber or user, unless

(a) the subscriber or user has given their consent to that use, and

(b) the subscriber or user has been provided with clear and comprehensive information in accordance with the Data Protection Acts which—

(i) is both prominently displayed and easily accessible, and

(ii) includes, without limitation, the purposes of the processing of the information.

Regulation 5(4) requires that the methods of giving information and consent should be as user-friendly as possible. Regulation 5(5) acknowledges the occasions where technical requirements may necessitate access to and technical storage of information in order to provide a service 'explicitly requested' by a user and stipulates that such storage and access must be 'strictly necessary'.

The evaluation and analysis of homecare companies' Cookie Policies and practices will be based on the requirements as set out in Regulation 5(3), 5(4) and 5(5), and must be analysed in conjunction with the terms of the GDPR. The standard for consent placed on controllers by the GDPR requires that it be obtained by means of a clear, affirmative act and be freely given, specific, informed and unambiguous.

Recital 24 of the e-Privacy Directive 2002/58 further clarifies the obligation to ensure the confidentiality of communications.

"*Terminal equipment of users of electronic communications networks and any information stored on such equipment are part of the private sphere of the users requiring protection under the European Convention for the Protection of Human Rights and Fundamental Freedoms.*

So-called spyware, web bugs, hidden identifiers and other similar devices can enter the user's terminal without their knowledge in order to gain access to information, to store hidden information or to trace the activities of the user and may seriously intrude upon the privacy of these users. The use of such devices should be allowed only for legitimate purposes, with the knowledge of the users concerned."

This last line reinforces and clarifies the requirement to establish active consent by the user as proof that they have 'knowledge' of the cookies that will be placed on their machine once they give their consent. The Court of Justice of the European Union (CJEU) Planet49 case [7] further clarified the nature and requirements around consent and established that the consent for the placement of cookies is not valid if it is obtained by way of pre-checked boxes which users must deselect to refuse their consent. In setting the use of cookies, the data controller normally needs the user's consent to use these types of technologies.[2] The data controller also needs to provide the user with certain prominently displayed, easily accessible, clear and comprehensive information on the technology being used and the purpose for which it is being used.[3] The position of the Irish DPC is stated clearly in their published report [8], that *"Users must be provided with easily accessible, 'clear and comprehensive' information on:*

- *The technology used by the website to collect personal data*
- *The purpose for which the collected data will be used."*

Article 12(1) of the GDPR stipulates that 'Information relating to processing must be presented in a concise, transparent, intelligible and easily accessible form, using clear and plain language. The information shall be provided in writing, or by other means, including, where appropriate, by electronic means.'

Therefore, when a user accesses a homecare company's website, they should be asked for their consent to 'cookies' in a manner which is clear and unambiguous and provides all the information necessary for them to make an 'informed' decision as to how they will proceed (e.g. give full or limited consent or refuse consent).

The Article 29 Working Party (WP29) has noted the practical problems related to obtaining consent, particularly if consent is necessary every time a cookie is read for the purposes of delivering targeted advertising [9]. WP29 also recommended limiting the scope of the consent to a period of time e.g. 1 year.

The practice of bundling cookie types by purpose, to make it easier (quicker) for the user to decide which cookies they are willing to accept, has become commonplace. The Planet49 judgment provided clarity on the practice of bundling of cookies when achieving consent, whilst allowing that consent 'does not need to be given for each cookie, but rather for each purpose. Where a cookie has more than one purpose requiring consent, it must be obtained for all of those purposes separately'. The DPC reports that in their sweep they found potential controller compliance issues which included "*the setting of cookies on landing without any engagement by the user with consent banners or other tools, lack of choice for users to reject all cookies and the bundling of consent for all purposes and the possible misclassification of cookies as 'necessary' or strictly necessary*". [8]

[2] Regulation 5(3)(a), ePrivacy Regulations 2011

[3] Regulation 5(3)(b)(i), ePrivacy Regulations 2011

B. Requirement for Speed of Access

Shastri et al [10] identified the phenomenon of 'metadata explosion' and the impact it has had in creating new workloads in relation to the GDPR. The research found that this explosion of metadata required changes to electronic storage systems to comply with the GDPR and resulted in significant performance overheads for database systems. In an increasingly digitised society, where speed is linked to performance, any reduction in the speed of database systems and their associated apps is not welcomed. This could potentially result in practices designed to speed up access to websites, databases or health apps taking precedence over requirements for privacy protection. The resulting delay in accessing a website due to the requirement placed on data controllers and website providers to provide the information necessary to achieve the 'informed' consent required to proceed and satisfy GDPR requirements is most likely unwelcomed by homecare company, website provider and client. It is the data controller's responsibility to ensure compliance with all relevant legislation and so they must provide sufficient information and in an appropriate format to facilitate the website users informed decision making.

Mulder [11] states in her study that the average length of the Privacy Policies analysed was 3,783 words (the largest was 11,344 and the shortest 347). Knowing that the average person reads 200-250 words per minute, it can be calculated how long it takes an average person to read a provided Privacy Policy prior to making an 'informed' decision. In Mulder's research this was 15-20 minutes.

III. Data Protection as a Service

Data protection by default from a technical perspective could be achieved in terms of cookie practices, as current software and programming can ensure no cookies are delivered unless informed consent is actively given and as such compliance is achieved.

Article 4 of the GDPR defines consent as *"... any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her".*

Article 7 determines that the data controller needs to be able to demonstrate that the data subject has given their consent. As a result, a written statement with an 'I agree' button combined with a Privacy Policy is one of the most common mechanisms used on websites to comply with GDPR.

The quality of the information provided in the written statement to inform such consent presents greater challenges when evaluating compliance. This author suggests such information should be standardised i.e. standardised acceptable information for each cookie purpose should be published by the DPC for inclusion on all cookie consent management platforms. Barrett [12] identified the influential role that the DPC is likely to play in evolving data protection legislation, not just in Europe but across the wider world due to its enforcement actions and guidance. Provision of standardised acceptable (to the DPC) content for inclusion on CMP's etc would be most effective guidance.

Mulder [11] suggests that the European Data Protection Board (EDPB) should, in cooperation with those involved in healthcare, create solutions to help ensure GDPR compliance. If the Data Protection Authorities (DPA's) are the masters of privacy protection, and as such fully understanding of the complex legislation, then perhaps the role of the EDPB and DPAs should evolve to that of providers of Data Privacy as a Service (DPaaS).

In the DPC's *Annual Report 2020*,[13] the Data Protection Commissioner identifies a requirement for the DPC role to expand '*to the benefit of organisations and data subjects alike including codes of conduct and certification*'. El-Gazzar and Stendal [14] identified the threat to privacy protection posed by emerging technology because of the legal framework's inability to keep up with technological advances, such that innovative technologies can inadvertently threaten the privacy of individuals. This author would suggest developing El-Gazzar and Stendal's approach to achieving improved GDPR compliance by providing 'governance and monitoring in context' through the provision of acceptable context specific content by the offices of the DPC for inclusion on CMP's.

IV. Methodology

This explorative research provides a better understanding of the problem of GDPR compliance pertaining to private homecare companies and their websites use of cookies. This research uses the content of homecare company websites, specifically their cookie practices (as experienced upon entering their websites) to evaluate if they are GDPR compliant.

A. Homecare Company Identification

Sixty-one homecare companies were identified using the following three sources:

1. The HSE published list of HSE 'approved' homecare support providers, where such homecare companies were franchisees the website of their main office located in Ireland was evaluated for content. (available on www.hse.ie)

2. Home and Community Care Ireland (HCCI) is the national representative body for homecare companies in Ireland. In 2020 Bedenik, the HCCI's Research and Policy Officer, published research that included the details of eighteen HCCI member homecare companies that were involved in the research [15]. This study included those same eighteen homecare companies identified in Bedenik's published research in its survey population.

3. An Internet search using the term "homecare Ireland" was initiated to find any homecare companies not identified at 1 and 2. Any additional homecare companies identified via this search also had their websites evaluated for content. Six of the companies identified did not have a website and so were discarded from the study. Each of the 55 websites were accessed using a cookie cleaned browser and their cookie practices identified, recorded, and analysed for GDPR compliance.

B. Cookie Compliance requirements

For a homecare company to be GDPR and e-Privacy compliant its website cookie practices must comply specifically with the Regulation 5(3) and 5(4) of the e-Privacy Directive. The following six questions based on the requirements set out in Regulation 5(3) and 5(4) were used to identify if such compliance was present.

1. Was a cookie pop up with choice presented when entering website?
2. Were there Cookie accept/reject/choose options presented?
3. Was cookie information as provided Clear & intelligible?
4. Was cookie information concise?
5. Were the cookie choice button/s presented in a balanced manner with equal prominence?
6. Were the website cookie Practices transparent and honest? i.e. Did they deliver cookies prior to permission being given and or having been specifically denied permission?

V. Results

The Chart below displays the consolidated results of the observed cookie practices of the homecare companies based on the criteria required to be compliant with the GDPR and e-Privacy directive.

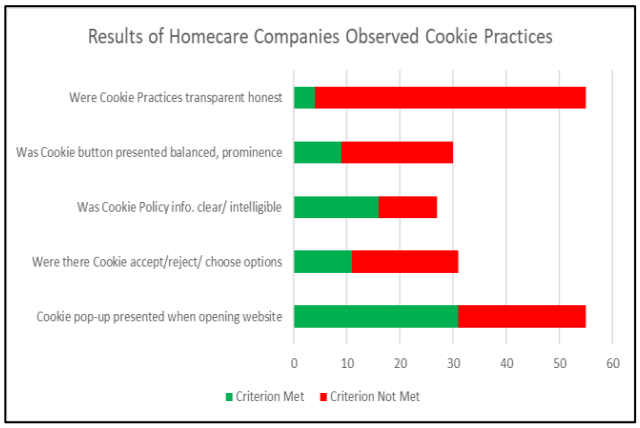


Chart 1. Consolidated Cookie Practice Results

49 of the 55 websites placed cookies prior to receiving permission, one of the companies placed 23 cookies without permission. Three of the companies had a default setting such that if accept was clicked you agreed to accept non- necessary cookies, this fact was hidden from the user.

Evaluating conciseness is problematic, as its subjective in nature, the average wordcount for those websites that published website specific privacy policies (WSPP) was 408 words (smallest was 36 words and largest was 1282 words). Is it possible to provide sufficient information in 36 words to inform a decision, is it realistic to think that a visitor will spend between two and six minutes reading the information before clicking the agree/disagree button?

Content of Homecare Company Website Privacy Policies

The homecare companies' website polices were of a particularly poor standard in terms of containing the required criterion set out in Articles 12 and 13 of the GDPR. Some of the content of the Privacy Policies viewed was of particular concern. Table 1 below displays direct quotations taken from the privacy policies and statements of homecare companies and are provided as further indicators of legislative breaches poor standards and lack of understanding of the requirements of GDPR and e-Privacy legislation as encountered by the author.

TABLE1 Examples Privacy Policy Statements

Policy Statements
"no information is collected that could be used by us to personally identity website visitors." *The next paragraph of the same policy acknowledges that the website cookie management platform (CMP) collects the IP addresses of visitors.*
"By visiting the website, you are accepting the terms."
By using our website you agree to us placing this type of cookie on your device," which included unnecessary cookies.
"By using our website you agree to the terms of the privacy policy in effect at that time."
"We may combine your account information with information we gather from your cookies."
"If you have signed up to marketing communications you have agreed to the transfer of your information to our email provider in the USA.". The same homecare company goes on to say it "does not generally transfer personal information abroad".
"A fee will be charged for access requests".

One homecare company used a different format when explaining and categorising cookie type and purpose in their Privacy Policy to that presented by their CMP, which also had a default setting such that the accepting 'non-necessary' cookies box was prechecked.

Aggregated cookie results

Four of the homecare companies could be considered as GDPR cookie practice compliant as there were no cookies detected on the author's computer having accessed their websites. None of the four published Cookie Policies or provided banners for evaluation etc. whilst still allowing access to their websites. Forty-four per cent of the sites accessed did not present cookie pop up banners, of the 31 companies that did only 35 percent (11) of those banners provided choices to reject accept etc that could be considered as GDPR compliant. Eighty-nine percent of the sites delivered cookies immediately upon landing on their site without any engagement by the user with consent banners or other CMP tools. Whilst some of these cookies may be considered strictly necessary or first party analytical, and perhaps pose little risk to those accessing the website, legislation requires consent and all CMP's, should be configured in such a way as to ensure legislative compliance.

The DPC [8] graded the 38 respondents to their cookie sweep using the 'traffic light' system: red, amber and green. For comparison purposes the same criteria is used to rank the cookie practices of the 55 homecare companies in this study.

GREEN: substantially compliant, any concerns straightforward and easily remedied.

AMBER: but at least one serious concern.

RED: with several serious concerns.

For comparison purposes the results were converted to percentages and are displayed in Table 1. As can be seen the DPC gave a range which facilitated those sites that were considered borderline between two grades.

Table 1. Comparing multi-sector cookie compliance as reported by the DPC to that of the homecare companies

Rating	Green	Amber	Red
DPC	5%-7%	53%-60%	32%-39%
This research	7%-29%	24%-33%	38%-47%

Four homecare companies were graded green and twelve were borderline amber to green, as the latter had minor issues that could be easily remedied, such as an unbalanced (in terms of prominence) banner button. Thirteen homecare companies were given an amber rating for having one significant concern, such as no cookie management information or banner displayed when entering the site. Five were given an amber to red as a result of having more than one but less than three concerns. Twenty-one homecare companies were given a red rating due to having several serious concerns including no banner or cookie information, no option to accept or reject cookies and the delivery of cookies against the expressed wishes of the site visitor. Five were given amber to red rating due to the nature and number of the concerns encountered.

VI. Discussion

The observed lack of compliance of companies that included the details of their Data Protection Officer (DPO) raises the question of DPO competency and is worthy of consideration. How they are selected, trained, and supported this author suggests is essential to achieving compliance. Currently the DPC requires that they are notified of the DPO's contact details (not all organisations are required to have a DPO). The matter of education and training is left to the discretion of the companies employing a DPO.

The DPC's published guidance on the qualifications for DPOs is taken from Article 37-39 of the GDPR and the Article 29 Working Party Guidance and is the further endorsed by the EDPB. It lists the relevant necessary skills and expertise required of a DPO as:

- Expertise in national and European data protection laws and practices including an in-depth understanding of the GDPR.
- In-depth understanding of how their organisation processes personal data.
- Understanding of information technologies and data security.
- Thorough knowledge of their organisation and the business sector in which it operates.
- Ability to promote a data protection culture within the organisation (DPC, Qualifications for DPOs).

Eighteen homecare companies provided the contact details of their DPOs, it would be reasonable to expect that at least eighteen of the companies would be fully compliant in terms of cookie practices (which was not the case), as DPOs are expected to have the relevant expertise skills and knowledge as outlined above.

This author would suggest the DPC should have the responsibility for certification of DPOs. They should evaluate and certify that a DPO has achieved and or displayed the necessary expertise and competencies to be awarded the title of DPO.

Grading the results of the homecare companies using a traffic light system facilitated comparison with the results of the DPC's research DPC [8]. However, it is the author's contention that homecare companies are adjudged to be either 'compliant' or 'non-complaint', whilst degrees of compliance (Red Amber or Green) may be considered when issues such as quantifying fines etc, the substantive question is and should be, is the homecare company compliant or not?

The subjective nature required to decide which breaches are more significant or which cookies are more wrong is immaterial if in the eye of the e-Privacy and GDPR legislation website owners must achieve informed consent before launching a cookie onto a visitor's computer. Issues regarding consent and implied consent have been clarified by both the GDPR and e-Privacy legislation and it is unacceptable to consider consent as being given just by landing on a website. Eighty-four per cent of the homecare companies did not provide an option to decline all cookies and dispatched cookies to the author's computer prior to the author responding to the banner or CMP.

As can be seen in the chart above the majority (strict interpretation would suggest 93%) of homecare companies were considered to be non-compliant.

Article 4 of the GDPR requires "*Clear and affirmative action*" on the part of the data subject to indicate consent.

Article 7 requires that the data controller be able to demonstrate that the consent has been given (usually by pressing an 'I agree' type button).

Regulation 5(3) of the e-Privacy Regulations requires consent to be "*explicitly requested"*.

Regulation 5(4) require that the information be provided in a user-friendly fashion and similarly the ability to consent and conversely reject cookies should be in a user-friendly fashion.

Regulation 5(5) acknowledges the occasion where technical requirements may necessitate access to and technical storage of information in order to provide a service "*explicitly requested"* by a user and stipulates that such storage and access must be "s*trictly necessary*". Evidence of the key words of 'explicitly requested' and 'strictly necessary' was noticeably absent in the numerous poor cookie practices as observed and reported above.

The CJEU Planet49 [7] case referred to earlier specifically excludes the use of prechecked boxes for unnecessary cookies. Again, this practice was observed and reported above. The legislation requires that all websites using cookies are legally compliant; whilst acknowledging that some cookies are of higher risk in terms of right to privacy. Cookie management platforms have the potential to provide the technical ability to ensure such risks are removed and begs the questions; Why do websites use CMPs that are not fit for purpose and why are all CMP's not fit for purpose?

Mulder [11] calculated that the average time required to read a privacy policy was 15 to 20 minutes, allowing for a reading rate of 200-250 words per minute. In our increasingly digitised society speed of access is considered essential. Will a website user spend sufficient time to read the content as provided on the CMP? If 'consent' is to be considered informed clear and concise, to satisfy legislative requirements? How much information (word count) its content (ease of understanding) and formatting are subjective risks, managed and owned by data controllers using CMP's. The competition for speed and ease of access versus legislative compliance.

Should content be sacrificed for speed? The author suspects that no DPA or legislator could or would ever agree to do so and that a potential solution would be to remove the necessity for data controllers to achieve consent through a technological solution as follows; ensure only necessary information is gathered, that cannot be further processed than that for which it was immediately required and that it is immediately anonymised and or erased so that the data subject's privacy is guaranteed.

The question as to why all websites do not use CMPs that are fit for purpose and protect users' privacy is relevant and worthy of further research. The following could be considered contributing factors:

- Legacy/older websites and CMPs not being maintained or upgraded in line with legislative requirements
- Ignorance of the legislative requirements and the responsibilities placed on data controllers and processors, inferred from the observation that 66% of homecare companies made no reference to 'recipients' (a GDPR requirement) in their Privacy Policies.
- Technical ignorance on the part of homecare companies with regards to website and cookie technologies
- Legislative ignorance on the part of technology providers. As CompTia [5] reports technical SMEs are lagging in awareness of the importance and responsibilities regarding data protection.
- Delayed enforcement actions on the part of the Irish DPC, although as can be seen from the DPC's 2020 activity this is something that has recently appeared on their 'action list'.

This author believes that there is a viable technical solution, provided by using websites and CMP's that are GDPR and e-Privacy compliant, requiring certification and or approved or by the DPC. This would create a situation whereby privacy is established by default in the manner decisions are presented to the user i.e., only strictly necessary cookies are let through unless the user actively adjusts the settings presented by the CMP having been presented with the relevant information in a manner acceptable to the DPC. The information necessary to personalise the templated CMP are provided by the homecare companies in a similar fashion to that proposed for the construction of Privacy Policies. All website providers and owners should be required to use a valid (DPA Certified) CMP and held culpable for failing to do so if they use cookies.

Website and CMP providers would be required to take part in a certification process that would allow them to market their products with the DPC's stamp of approval. Enabling perhaps the less versed (in terms of IT and data protection) homecare providers to be comforted that the product they invest in is meeting the required standards.

VII Conclusions

The homecare companies evaluated in this research were found to be lacking in terms of their compliance with the e-Privacy Directive and the GDPR. The homecare sector is a vital part of the social fabric of Ireland, providing essential services to the most vulnerable in our society. To do so, they are trusted to provide care and assistance in an appropriate manner, often unsupervised in the homes of the vulnerable. It is essential that homecare companies preserve the dignity and privacy of those they are entrusted to care for. Homecare companies use privacy and trust in their marketing strategies as they understand their importance to those in need of care. It would be hugely detrimental for any homecare company to be publicly 'outed' as being in breach of legislation specifically designed to protect an individual's right to privacy. The marginal gains in terms of marketing strategies and data analytics (resulting from non-consented cookies landing) are far outweighed by the impact and reputational damage that would be caused to a homecare company found guilty of breaching data privacy legislation. To that end this author suspects the breaches and poor practices identified by this research are more likely the result of ignorance of technology and the applicable legislation. Ignorance of the law, however, is never an acceptable excuse.

If the necessity to use cookies continues then to move cookie practices to green from red (Table 1 refers) there are broadly speaking two elements required: a technical element and a content/information-based element. The technical element is easily achieved and requires that all CMP's should be DPC approved/certified with the following automated settings.

1. No cookie should be installed on a user's computer prior to receiving consent, the CMP default option should be set to 'reject all'
2. The CMP should provide information on cookie types utilised that is approved/certified by DPC
3. All CMPs should include options to
    a. Reject all cookies (always the prechecked default option)
    b. Accept necessary cookies only
    c. Accept other 'Stated' categories of cookies (with a suitable certified explanation of the function and purpose of each category)
4. All cookies should have an appropriate expiration date.
5. The options buttons should be balanced in terms of prominence (size and colour).

The content or information element is essential to ensure that consent is considered as properly informed, this should be in a standardised format explaining the category and function of the cookies utilised in a format (clear and concise) and approved/certified by the DPC. An alternatively is not to use

cookies or ensure automated anonymisation is achieved and no identifiers are captured to protect the privacy of website users. As mentioned earlier the evolution of the DPC's role to that of a service provider would help to improve compliance and protect privacy. The author agrees with Mulder's recommendations [11] (2019:19) that the EDPB and its DPAs become involved in "creating solutions" and with El-Gazzar and Stendal's suggestion [14] (2020:270) that GDPR compliance by design cannot be achieved without close cooperation and collaboration. Similarly, Yang et al [5] identifies the requirement for coordination of people and policies to achieve effective governance further support for these authors suggestions that DPAs become 'providers of acceptable content' for users. The suggested enhanced role of Ireland's DPC should include the following:

- Provide sector specific Privacy Policies and information content using appropriate software to enable personalisation as required
- Provide certification/approval for website and CMP providers

As the levels of noncompliance with both GDPR and the e-Privacy directive were so high in terms of homecare companies' cookie practices. The homecare companies DPO's could be considered as failing to display the skills and expertise required of that role as published by both the EDPB and DPC, specifically in terms of understanding information technologies, data security and the regulatory environment. The latter considered essential by CompTia [5] as a function to ensuring cybersecurity.

This research identified compliance failure in terms of homecare companies' websites use of cookies, performance failure on the part of homecare companies' DPO's. And while it may be inadvertent and unintentional it poses a risk of reputational and financial damage to homecare companies. The failings are more likely to have resulted from a lack of understanding on the part of website and CMP providers of the legislative environment. Evidenced by the fact that the technical solution is easily achieved. The research also identified an opportunity for the DPC to evolve their role to become certification providers and providers of approved content to resolve or reduce the negative impact operating in a complex and evolving legislative environment causes on data protection compliance.

VII Recommendations

Future Research should engage with homecare companies and website and CMP providers to establish their level of awareness of the relevant legislation.

An analysis of the effectiveness of DPO's should be conducted in terms of their ability to aid compliance and the potential influence that a DPC controlled certification process for DPO's would have on improving compliance.

VIII References

[1] Health Service Executive National Service Plan 2020, 2021.Available at: www.hse.ie/eng/services/publications/national-service-plan-2020.pdf . Accessed 01 March 2021

[2] S.I. No. 336/2011 European Communities (Electronic Communication Networks and Services) (Privacy and Electronic Communications) Regulations 2011. Available https://www.irishstatutebook.ie/eli/2011/si/336/made/en/print?q=336/2011

[3] General Data Protections Regulation, 2016 (GDPR) EUR-Lex - 02016R0679-20160504 - EN - EUR-Lex (europa.eu)

[4] Yang, L., Li, J., Elisa, N., Prickett, T. and Chao, F., 2019. Towards big data governance in cybersecurity. *Data-Enabled Discovery and Applications*, *3*(1), pp.1-12.

[5] Comptiacdn.azureedge.net. 2021. [online] Available at: https://comptiacdn.azureedge.net/webcontent/docs/default-source/research-reports/research-report---state-of-cybersecurity-2020.pdf [Accessed 05 March 2021].

[6] Data Protection Act 2018. Available: https://www.irishstatutebook.ie/eli/2018/act/7/enacted/en/html?q=data+protection+act+

[7] "Planet49: Reference for a preliminary ruling", 2019, Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=218462&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=2294151

[8] "Data Protection Commission (Ireland), 2020a. Report by the Data Protection Commission on the use of cookies and other tracking technologies.

[9] "Article 29 Data Protection Working Party, 2010. Opinion 2/2010 on online behavioural advertising" [Online]. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2010/wp171_en.pdf . (Access Date: 27-03-2020)

[10] Shastri, S., Banakar, V., Wasserman, M., Kumar, A. and Chidambaram, V., 2020. Understanding and benchmarking the impact of GDPR on database systems. *Proceedings of the VLDB Endowment*, 13(7), pp.1064-107

[11] Mulder, T., 2019. Health Apps, their Privacy Policies and the GDPR. European Journal of Law and Technology, 10 (1 2019).

[12] Barrett, C., 2020. Emerging Trends from the First Year of EU GDPR Enforcement. Scitech Lawyer, 16(3), pp.22-35.

[13] Data Protection Commission (Ireland), Annual Report 2020. Available https://www.dataprotection.ie/en/news-media/press-releases/data-protection-commission-publishes-2020-annual-report. Accessed 06 March 2021

[14] El-Gazzar and Stendal., 2020. Examining How GDPR Challenges Emerging Technologies. *Journal of Information Policy*, 10, p.237.

[15] Bedenik, T., 2020. 'An Inquiry into the Lived Experience of Covid-19 in the Home care Sector in Ireland: Experiences of Home Care Provider Organisations', Home and Community Care Ireland, Available at: https://hcci.ie/covid-19-provider-research/

The Critical Success Factors for Security Education, Training and Awareness (SETA) Programmes

Areej Alyami
Business Information Systems
Cork University Business School
University College Cork
Cork, Ireland
a.alyami1988@gmail.com

David Sammon
Bussiness Information Systems
Cork University Business School
University College Cork
Cork, Ireland
dsammon@ucc.ie

Karen Neville
Business Information Systems
Cork University Business School
University College Cork
Cork, Ireland
KarenNeville@ucc.ie

Carolanne Mahony
Business Information Systems
Cork University Business School
University College Cork
Cork, Ireland
carolanne.mahony@ucc.ie

***Abstract* —This study explores the Critical Success Factors (CSFs) for Security Education, Training and Awareness (SETA) programmes. Data is gathered from 20 key informants (using semi-structured interviews) from various geographic locations including the Gulf nations, Middle East, USA, UK, and Ireland. The analysis of these key informant interviews produces eleven CSFs for SETA programmes. These CSFs are mapped along the phases of a SETA programme lifecycle (design, development, implementation, evaluation).**

Keywords: SETA; Security; CSFs; Key Informant

I. INTRODUCTION

One of the most vital and prominent approaches to managing IS security risks and safeguarding IS and information assets in an organization is its Security Education, Training, and Awareness (SETA) programme. Many researchers recommend establishing a SETA programme as part of the organization's overall security strategy (Alshaikh et al, 2018; Kirova and Baumöl, 2018; Tsohou et al., 2015; D'Arcy et al., 2009). In the literature, the SETA programme is also referred to as IS security training (Parrish and San Nicolas, 2012; Karjalainen and Siponen; Heikka, 2008), and an IS awareness programme (Bauer et al., 2017; Tsohou et al., 2015). Peltier, 2005). The importance of SETA programmes has received significant academic attention: various studies discuss the use of a SETA programme to improve employees' behaviour (Alshaikh et al., 2019; Bulgurcu et al., 2010; Mahmood et al., 2010), to comply with IS policy (Cram et al., 2019; Barlow et al., 2018; Puhakainen and Siponen, 2010), and to increase the level of awareness and reduce IS security risks (Tsohou et al., 2015; Karjalainen and Siponen, 2011; D'Arcy et al., 2009).

Despite the prominence of SETA programmes for organisational IS/cyber security governance *"only a small portion of practitioners"* claim that their SETA programmes are *"very effective"* (Hu *et al.,* 2021a, p.1). It is reported that poor SETA programme effectiveness is linked to the programmes failure to achieve its goal of impacting positively on employee security-related behaviours (Alshaikh *et al.,* 2021; Hu *et al.,* 2021a; He and Zhang, 2019; Alshaikh *et al.,* 2019). A lack of a *"systematic understanding"* of the *"nature of SETA programmes"* and their impacts on *"security-related beliefs"* is viewed as a possible reason for this lack of effectiveness (Hu *et al.,* 2021a, p.1). In fact, Alshaikh *et al.* (2021, p.1) argue that existing SETA programmes are *"suboptimal"* as they *"aim to improve employee knowledge acquisition rather than behavior and belief"*. Therefore, more theorizing and conceptual clarity is needed in investigating the effectiveness of SETA programmes (c.f. Alshaikh *et al.,* 2021; Hu *et al.,* 2021b; Kirova and Baumöl, 2018; Puhakainen and Siponen, 2010). This paper sets out to address this research need by exploring the Critical Success Factors (CSFs) for SETA programmes.

The paper is organized as follows: Section 2 presents a background to SETA programmes. Section 3 describes the methodology: the data gathering and the data analysis techniques. Section 4 presents the findings: the CSFs for SETA programmes. Lastly, section 5 presents the conclusion and the plan for future research.

II. SETA PROGRAMME BACKGROUND

The Security Education, Training and Awareness (SETA) programme is an educational process designed to reduce the number of accidental security breaches that occur due to a lack of employee

awareness of IS security (Whitman and Mattord 2008; D'Arcy *et al.,* 2009; Puhakainen and Siponen, 2010; Han *et al.,*2017; Alshaikh *et al*., 2018; Barlow *et al.,* 2018; Yoo *et al*., 2018; Dhillon *et al*., 2020). The existing literature distinguishes between education, training, and awareness terminologies based on their specific aim and target. For example, Whitman and Mattord (2008) propose that the aim of 'education' is for security experts to gain a deep knowledge regarding the design and implementation of a SETA programme; 'training' helps employees to acquire a level of skill that enables them to perform their job securely; and 'awareness' encompasses the delivery of information and informal training to employees to increase their awareness of potential risks and IS security issues. Therefore, the significance of SETA programmes is widely accepted by both academics and practitioners (Wilson and Hash, 2003; D'Arcy *et al.*, 2009; Tsohou *et al*., 2015; Alshaikh *et al*., 2018). Based on a review of the literature, SETA programmes typically address the following:

1. provides employees with knowledge regarding organizational information threats and IS security (D'Arcy *et al*., 2009; Yoo, *et al*., 2018; Dhillon *et al*., 2020);
2. clarifies existing technical and procedural countermeasures available to employees (Pastor *et al*., 2010; Silic and Lowry, 2020);
3. determines the possible sanctions for security policy violations in the organization (Siponen and Vance, 2010; Karjalainen *et al.,* 2013; Herath, *et al*., 2018), and
4. improves employees' awareness of their roles and responsibilities in protecting the organization's information assets (D'Arcy *et al*., 2009; Lebek *et al*., 2014).

In fact, there is a stream of research that examines SETA programmes by focusing on an individual employee (micro-level) analysis and explores the factors that affect security behaviour directly or indirectly. This then allows exploration of the factors that influence security-compliant behaviour (Burns *et al*., 2015; Alshaikh *et al*., 2019). Another research stream focuses on the individual but also identifies organizational-level factors that influence information security compliance policies (Chen *et al*., 2015; Lowry *et al*., 2015; Burns *et al*., 2018). A third stream focuses on an organizational level (macro-level) analysis, providing directions for the design and implementation of awareness programmes, change of information security strategy, power relations, and allocation of responsibilities (Straub and Willke, 1998; Peltier, 2005; Puhakainen and Siponen, 2010; Karjalainen and Siponen, 2011; Tsohou *et al*.,2015).

However, research is still required on the design, development, implementation, and evaluation phases of SETA programmes (Alyami *et al.,* 2020; Alshaikh *et al.,* 2018). For example, where empirical studies investigating the effectiveness of SETA programmes exist, they fail to examine all phases of the SETA programme lifecycle (design, development, implementation, evaluation), tending to focus more on one or two of the lifecycle phases. For example, Puhakainen and Siponen (2010) propose a method to **design** an information security awareness programme, while Okenyi and Owens (2007) identify four factors that contribute to the **development** of a successful SETA programme. Furthermore, Silic and Lowry (2020) report on the use of an IT artefact (a gamified security training system) enabling a SETA programme **implementation**, while Rantos *et al.* (2012) provide a methodology to assist organisations in the **evaluation** of their awareness programme efforts.

Leveraging the SETA programme lifecycle phases (design, development, implementation, evaluation), we now explore the CSFs for SETA programmes. Each one of these CSFs is mapped to the relevant lifecycle phase. This mapping produces 11 CSFs for SETA programmes. In the next section, we present further details on our research methodology.

III. RESEARCH METHODLOGY

To fulfil the research objective, this research follows an exploratory design. As agreed by Marshall and Rossman (1989), the purpose of an exploratory research approach is to investigate a little-understood phenomenon. The CSFs for SETA programmes are the outcome of this exploratory research approach.

A. DATA GATHERING

In this research, we adopt the "key informant" approach for data gathering and engage with key informants through semi-structured interviews. A key informant is an expert in a particular field who is highly experienced and knowledgeable. According to Marshall (1996), the five criteria for selecting a key informant are as follows: (1) knowledge (the informant should have a depth of information and experience of the phenomenon); (2) willingness (the informant must be willing to communicate and share their knowledge and experience); (3) communicability (the informant should be able to transfer their knowledge in a way that is understandable to the interviewer); (4) impartiality (the informant should be unbiased, and any relevant biases must be disclosed beforehand to the interviewer); (5) role in community (the informant should understand how their role contributes to an understanding of the phenomenon). Therefore, key informants were selected based on their position, experience, and professional knowledge about IS/cyber security, particularly SETA programmes.

Interviews are one of the most suitable techniques for gathering valuable data from experts (Marshall and Rossman, 1989). The semi-structured interview is suited to exploring new ideas, capturing new phenomena, and identifying the rich contextualized detail of complex concepts. Twenty individual semi-structured interviews were conducted with selected key informants from various geographic locations which included the Gulf nations (Saudi Arabia, United Arab Emirates, Qatar and Kuwait), the Middle East (Egypt and Lebanon), USA, UK and Ireland. Table 1 provides a list of the key informants' positions, years of experience and interview duration.

Table 1. The key informants' positions, years of experience, and interview duration

Key Informant No.	Country	Role	Experience (years)	Interview duration (minutes)
1	Saudi Arabia	IS security consultant	more than 12 years	60
2	Saudi Arabia	CISO (chief information officer)	almost 8 years	45
3	Saudi Arabia	Supervisor in the cybersecurity department	10 years	55
4	Kuwait	Cyber security leader	almost 22 years	60
5	Lebanon	Governance and risk management compliance manager	10 years	40
6	Qatar	Senior manager for governance risk and compliance	12 years	45
7	UAE	InfoSec training lead	10 years	40
8	UAE	Consultant in IS security	more than 17 years	50
9	Saudi Arabia	CISO (chief information officer)	15 years	55
10	Kuwait	CISO (chief information officer	8 years	40
11	USA	Consultant in IS security	20 years	60
12	UK	CISO (chief information officer)	almost 20 years	55
13	USA	Director for cyber leadership and strategy solutions	25 years	45
14	Kuwait	Head of information security governance	20 years	50
15	Saudi Arabia	Cyber security consultant	10 years	60
16	Egypt	Head of cyber security	20 years	55
17	UK	Security Awareness Manager	15 years	50
18	USA	Director of Security Awareness	over 20 years	45
19	Ireland	Senior lecture in IS security	17 years	45
20	Ireland	IT security officer	21 years	50

All of the interviews started by introducing the objective of the research. Each interviewee was then asked to provide a brief summary of their background. Thereafter, topics relating to the CSFs for SETA programmes, throughout the lifecycle phases (design, development, implementation, evaluation), were discussed. The interviews were conducted in two languages, some in Arabic and some in English, and the Arabic interviews were translated into English also. All the interviews were transcribed line-by-line and checked against the voice recordings, where necessary, to ensure the accuracy of the transcription of the interviews.

B. DATA ANALYSIS

Data analysis is a crucial step in qualitative research (Leech and Onwuegbuzie, 2008). Its main purpose is to develop an understanding of the phenomenon of interest (Kawulich, 2004). In this research we adopted an inductive open coding approach as part of our qualitative data analysis. This coding technique is aimed at generating concepts from field data (Walsham, 2006) and according to Strauss and Corbin (1990, p.61) open coding is defined as *"the process of breaking down, examining, comparing, conceptualizing, and categorizing data"*. Moving through the open coding process afforded us the opportunity to identify the concepts or key ideas hidden within the key informant interview data and related to the phenomenon of interest (c.f. Bhattacherjee, 2012). As part of our open coding, we also grouped similar concepts into higher-order, more abstract concepts, called categories.

When all 20 key informant interviews were transcribed, the data analysis commenced using sentence-by-sentence coding to identify relevant

codes. The open coding procedure for the 20 key informant interviews resulted in 212 coded excerpts relating to the factors impacting on the effectiveness of a SETA programme. These 212 coded concepts led to the emergence of 15 categories mapped across the 4 SETA programme lifecycle phases. Specifically, the code/category distribution is as follows: **design** phase – 95 codes – 8 categories; **development** phase – 27 codes – 4 categories; **implementation** phase – 50 codes – 5 categories; **evaluation** phase – 40 codes – 3 categories. Thereafter, unpacking the categories with at least five key informant voices (25% coverage) led to the emergence of the 11 CSFs for SETA programmes. The next section discusses the research findings.

IV. FINDINGS: THE CSFs FOR SETA PROGRAMMES

Critical Success Factors (CSFs) are defined as *"key areas where things must go right in order to successfully achieve objectives and goal"* (Bullen and Rockart, 1981, p.9). CSFs have been widely researched, debated and cited across a wide range of information systems (IS) topics, which accounts for their continuing popularity. In essence, their simplicity, as a statement of focus and action, is their most valued characteristic. Given the purpose of this study, the remaining sections present the CSFs for SETA programmes.

1. CSF#1: Conduct an Initial Assessment of Employee Security Awareness

This CSF highlights the fact that conducting an initial assessment is an essential factor in designing a SETA programme. Primarily, a focus on determining what the employees understand about the organization's security policy is crucial, along with an understanding of their appreciation of the risks associated with current cyber security threats. Within this study, key informants suggest conducting an initial assessment using tools like surveys or quizzes in an effort to gauge how knowledgeable the employees are about IS security issues. For example, one key informant mentions *"completing a test on IS security to realize what the employee understands exactly about information security"* while another informant suggests "*an initial assessment to understand what is working and what is not working".* It is also noteworthy that employees at various levels within the organisation will have different types of assessments to complete. For example, the assessment that an IS security manager completes will be different to the one completed by the end-user. As noted by one of the key informants: *"each level has a specific security awareness programme regarding cybersecurity".* Therefore, this CSF emphasizes that identifying the current level of understanding around cybersecurity issues, as part of the design phase of a SETA programme lifecycle, will increase the likelihood of successful SETA programme outcomes.

In comparing these findings with those presented in the literature, a number of observations can be made. Several studies have called out the importance of understanding the need to establish a SETA programme and identify the security awareness plan that addresses employee needs (Alshaikh et al., 2018; Puhakainen and Siponen, 2010; Vroom and von Solms, 2002). In fact, Peltier (2005) suggests that when organizations use assessments to determine what the expected threats are and what the associated risk level of these threats is, then the information needed to protect the organization is provided. The outcome of the assessments helps to determine the needs that must be covered. This kind of assessment assists in designing an appropriate SETA programme and makes it easier to prioritize the design to meet a specific need (Okenyi and Owens, 2007). As a result, this step is crucial to show the current position of the organization with regard to security reports, previous incident attacks and previous threat responses.

2. CSF#2: Build Security Awareness Campaigns

This CSF highlights the fact that targeted awareness campaigns can update employees (or end-users) on how to mitigate against the potential risks associated with an IS security threat and keep them informed on what is coming, and most crucially, why they need to care. Within this research study, key informants state the need for discussion at the end of an IS security training session or awareness campaign. It is as part of these conversations that individuals understand the security awareness message. For example, one key informant noted: *"what is important in this session is to assess if the people are actually getting your security message…".* In addition, a security awareness campaign should be rolled out every three months and a follow-up also organized with employees, for consistency and reliability, and to emphasize the importance of the security awareness programme to the organization. As stated by another informant: *"to build a security awareness and training program, you need to communicate with all the stakeholders and say this is coming. This is why you care. People need to understand why it is important…".* Therefore, to build a security awareness campaign that plays an important role in the success of a SETA programme is of critical importance.

In comparing these findings with those presented in the literature, a number of observations can be made around the criticality of building a security awareness campaign as part of a SETA programme. For example, Rantos et al., (2012) discuss launching the awareness

campaign across the company, to cover all IS security topics, as a vital element of measuring the effectiveness of the SETA programme. Several studies highlight the need to design an awareness campaign, as a periodic short communication, to clarify the importance of the SETA programme in terms of protecting the IS assets, personal data, enhancing IS security awareness, complying with IS security policy, and reducing IS security risks (Vroom and von Solms, 2002; Puhakainen and Siponen, 2010). Therefore, formal awareness campaigns are communications with employees with the specific aim of: [1] increasing the understanding of, and [2] reducing the likelihood of, harmful information security practices within the organization (D'arcy et al., 2009; Hearth et al., 2018).

3. CSF#3: Design for Cultural Context and Employee Cultural Diversity

This CSF focuses on the criticality of understanding the cultural diversity in the organization when designing a SETA programme, simply because the cybersecurity message can be interpreted differently from one culture to another. Employees come from different backgrounds, and it is necessary to understand this diversity. Various aspects of cultural context require focus when designing a SETA programme, such as: language, knowledge, level of education, age, and gender. All these aspects contribute to a successful SETA programme outcome. For example, within this research study, the key informants come from many countries and all these countries have their own culture. Therefore, if our key informants represented a typical organisation's employees, then these differentiations would need to be considered when designing a SETA programme. For example, the cultures of Saudi Arabia, Egypt, and UAE care more about language, and as a result use artefacts for SETA programmes, such as videos and posters in Arabic, to make the message more attractive and easier to understand. As stated by one key informant: *"culture is an important factor to consider when you want to design an awareness program, we design the videos in the Arabic language that contains street language; we noticed the employees interact with these kinds of videos".* However, understanding culture across different geographical locations in terms of knowledge, language and education further contributes to the success of a SETA programme. As commented by key informant: *"...design the SETA programmes in a way that is close to the culture to make it a success."* Therefore, each culture has specific characteristics that make it unique from other cultures and this must be appreciated to ensure the effectiveness of the SETA programme.

In comparing these findings with those presented in the literature, a number of observations can be made. Previous studies address 'culture' in the context of IS security practice. For example, Hovav and D'Arcy (2012) examine the influence of the culture on the IS security policies, training, and monitoring. In fact, to understand culture in terms of IS security practice is to understand individual differences within each cultural context (c.f. Walsham, 2002). These cultural differences can be beliefs, norms, and values in a social setting, known collectively as a country. Thus, different cultures require different IS security interventions (Kirova and Baumöl et al., 2018; Karjalainen et al., 2013; Von Solms and Von Solms, 2004). Thus, understanding the cultural context is an essential factor when designing a successful SETA programme.

4. CSF#4: Make a Yearly Plan to Align Goals and Objectives

This CSF highlights the importance of communicating the SETA programme objectives (knowing what is required to be delivered) clearly and consistently to the employees. It is also important to ensure that the SETA programme goals meet the specific needs of the organization (as captured in its strategy) and these two aspects are aligned during the design phase. Within this research study, key informants suggest that a yearly plan be devised to determine the objectives and design of the SETA programme based on the activities it wants to achieve. For example, one key informant states: "*...every year we make a plan, determine our goals or objectives of the year, then we design activities for the awareness programme to see how to execute the plan….".* In addition, each year, most organizations update their objectives regarding the SETA programme. Another key informant commented: *"...if it wasn't specifically designed, the organisational SETA programme would not succeed. As well, if its objectives are not associated with the strategies of the institution, it will not work".* This suggests that organisations should create a plan for designing a SETA programme and that plan should contain what is necessary to be delivered, such as the types of IS security issues or topics.

In comparing these findings with those presented in the literature, several observations can be made around tailoring SETA programmes to meet specific organizational needs. For example, Rantos and Manifavas (2012) discuss methods to create an effective awareness programme. One of those methods is based on planning around the specific needs (e.g. materials to cover on the security awareness programme) to meet the organization's goals. Other studies mentioned that identifying the objectives is the initial step when establishing a SETA programme (Peltier, 2005; Hansche, 2001). Most organizations

initiate the design of the SETA programme with specific goals in mind. For example, a plan and new security policies to address any ongoing challenges (from years previous) and to ensure the delivery of a successful SETA programme. Therefore, to establish the SETA programme, one must have a clear goal that supports the organization's overall mission.

5. CSF#5: Adhere to Organisational Security Policy and the "Law of the Land"

This CSF focuses on the guidelines and procedures needed to protect the IS assets of the organization. These factors can be regulation or legislation that help to modify employee IS security behaviour. It is critically important that all of the organizational security policies and the "law of the land" are adhered to when designing a SETA programme (e.g., General Data Protection Regulation (GDPR) in Ireland, and the Saudi Arabian Monetary Authority (SAMA) in Saudi Arabia). Within this research study, key informants stress that the organization should be aware of all regulations and policies. Each country has its own rules and regulations regarding data privacy and data security. As mentioned by one key informant: *"most of the organizations design SETA programmes in-house, and these programmes should align with their security policy. For example, laws in some countries are different".* In addition, all employees in the organization are obliged to be aware of the information security policy within their organization. Each organization has its own policies, for instance, the restriction on the sharing of passwords among employees and other social engineering issues. For example, one key informant stated: *"all members of the organization, from the board to the technical employee, have a duty to be aware of the information security policy and privacy".* Thus, understanding the business requirements and their policies are fundamental to designing a SETA programme.

In comparing these findings with current literature, a number of observations can be made. Some studies focus on the security policy and regulations in building a SETA programme (D'Arcy et al, 2009; Peltier, 2005). The security policies are presented to the employees to show what is expected from them. Therefore, to make a SETA programme successful, the employee should follow the policies and regulations in order to deal with issues such as: how to deal with suspicious sites; how to keep company data confidential; and which information can be shared on social media.

6. CSF#6: Know Your Audiences to Ensure Content Suitability

This CSF highlights the importance of allocating the appropriate privileges to employees, using their organizational role to determine their security responsibilities. Identifying "who your audiences are" is critical in designing a SETA programme to ensure content suitability. Within this research study, key informants explain how most organizations set up a SETA programme based on their audiences' levels. Therefore, materials used must be appropriate for each level to ensure that employees understand the contents of the security training. For example, one key informant comments: *"we start to plan to design a SETA programme based on audience classification, it's important to provide the material based on knowing those who we are speaking to understand what we are saying…".* It is clear that a top management employee has different security training to a new graduate employee. As one key informant states: *"so employees working in operation sites, oil production, or HR, etc., they might see some different pieces of training and sometimes different material".* Thus, each job role in the organization has specific responsibilities such that the requisite IS security training needs are different.

In comparing these findings with those presented in the literature, Pelter (2005) discusses establishing a security awareness programme by classifying the audience to ensure the security message is communicated effectively. Accordingly, a SETA programme must comprise a plan to transmit the IS security message to the target audience (De Maeyer, 2007; Siponen, 2000). It can be argued that identifying the target audiences in designing a SETA programme is the main step toward its success; thereby delivering particular security training, with appropriately suitable material, to each employee.

7. CSF#7: Sustained Communication of Relevant Messages

This CSF is based on how to communicate with audiences regularly and how to follow up with updated materials and topics. The security message should be repeated differently because the audience can lose concentration and forget. Thus, continuous communication with employees regarding IS security practices is an effective way to assist them in reducing security incidents and breaches. Within this research study, key informants highlight the importance of sustainable communication with the employees for the development of the SETA programme. For example, one key informant notes: "*we need to direct and inform the employees that this issue of security awareness is not only crucial in their work environment but also in their life routine".* Effective communication clarifies why some issues are not permitted. It can show the employees examples of real-life cases of human errors at play while informing them of the enormity of the problems by using pictures

and real stories. As stated by one key informant: *"…when we have a real human error, telling them this is a real problem by proving this with pictures and real stories with consequences, is invaluable…."*. In addition, security training and awareness materials must be updated based on current situations. For instance, one key informant comments: *"we are facing problems such as Covid-19 and working remotely. It is important to have materials based on this situation, so they can connect both things and will never forget whatever was given".* Thus, it is necessary to always remind the employees that IS security issues exist all the time, whether in the work environment or in one's personal life.

In comparing these findings with existing literature, we find a limited number of studies that examine the impact of communication on the effectiveness of a SETA programme. This presents an opportunity for further research. For example, Barlow et al. (2018) state that more research on the role of communication in delivering a SETA programme is required. Therefore, from a practical point of view, sustained communication plays an important role in the success of a SETA programme.

8. CSF#8: Apply Diverse Methods to Deliver Security Awareness Messages

This CSF highlights that organizations use various approaches to deliver SETA programme messaging. For example, they can deliver security awareness messages via SMS, emails, online courses, face-to-face meetings, videos, quizzes, and posters. In addition, by placing security awareness messages on internal screens in public areas, such as corridors, employees are reminded frequently of this security issue. Thus, organizations determine the best methods to use to implement their SETA programme messaging based on their resources, size, and budget. Within this research study, key informants identified the various methods to deliver a successful SETA programme. As commented by one key informant: *"the best security awareness programmes include various IS security delivery methods because we have to consider individuals' differences".* The popular method used to implement a SETA programme is computer-based training (CBT) that includes all training materials and quizzes. It is a platform that anyone can access anywhere. However, the latest trending method is 'gamification' which is a very interactive application like playing a game. The organization engages the user by sending out materials or videos, and employees can watch the videos and answer the questions accompanying them. For example, one key informant states: *"the new trend in Cybersecurity Awareness is 'gamification' - conducting games for employees…".* All organizations have access to this and other methods to promote security awareness to their employees.

In comparing these findings with those presented in the literature, several studies discuss different methods to implement a SETA programme (Silic and Lowry, 2020; Bauer et al., 2017; Tsohou et al., 2015; Johnson, 2006; Peltier, 2005). For example, Silic and Lowry (2020) present a study that aims to improve security training in organizations by applying a gamification approach. While other studies discuss different communication channels such as posters, videos, emails etc. to deliver a SETA programme (Johnson, 2006; Peltier, 2005). It can be argued that the successful implementation of a SETA programme can be determined by a diversity of delivery methods aligned with individual differences.

9. CSF#9: Motivate Employees to Engage in Security Awareness

This CSF highlights that employees can be encouraged to adhere to IS security policies by earning a bonus or other recognition (reward) based on their practices. This can have a positive impact on the effectiveness of the organization's SETA programme. In this research study, key informants mentioned several methods to motivate employees to embrace IS security training. For example, employees can be invited to complete several tasks such as quizzes or videos that are assigned scores. These scores can waive other requirements such as attending security awareness courses. This method was described by a key informant as follows: *"I think it is a really good incentive for employees. If the employee can pass the quiz with 100%. You don't have to watch the video…".* This type of motivation encourages the employee to learn necessary materials to pass quizzes. An employee can also be motivated by attending events or celebrations that promote the organization's security policy. One key informant from Saudi Arabia mentions that *"some government agencies contributed to arranging activities and are welcoming of the employees' families and their children by giving colouring books to their children…".* These events include recommendations about appropriate security practices to promote security awareness. Additionally, focusing on the social side motivates employees to attend the events and understand the IS security issues in a social setting.

For this study we use the definition of 'motivation' proposed by Rogers (1975), where motivation can be either intrinsic (doing something since one finds it interesting) or extrinsic (doing something since one is obliged to, or to be rewarded). Several studies examine the influence of motivation to sustain compliance with IS security policy (Puhakainen and Siponen, 2010; Herath and Rao, 2009), change employee

behaviour (Alshaikh et al., 2018; Kirova and Baumöl, 2018; Karjalainen et al., 2013) and reduce IS security risk (Zani et al., 2018). Although we did not find studies that examine the impact of motivational aspects on the effectiveness of SETA programmes, it is an area that requires further research.

10. CSF#10: Maintain Quarterly Evaluation of Employee Performance

This CSF focuses on providing a year-end evaluation summary to measure each employee's performance, level of awareness, and number of training sessions completed. This evaluation is a report of the employee's progress and provides guidance on improvements to be made. For example, one of the significant tools for evaluating employees' performance in the annual report is the Key Performance Indicators (KPIs) related to IS security issues, such as: cybersecurity attacks, phishing campaigns, sharing password policy breaches, etc. Each quarter, most organizations use KPIs to evaluate employee performance and the percentage that fulfil the training requirements, in order to assess the knowledge retained by employees and thereby review the effectiveness of the SETA programme. Within this research study, key informants highlight several techniques to assess the employees' responses to the SETA programme. One of the techniques used is a survey/questionnaire to evaluate employee knowledge before and after they have undergone training. This type of evaluation answers important questions such as: have we overcome the challenges?, or, did we make the same mistakes? As one key informant comments: "*...conducting a questionnaire before the training and after to know the amount of knowledge the employee is getting from the security context. Then we can measure the effectiveness of these programmes…".* Another technique is the use of quizzes. After completing IS security training, passing a quiz can be an effective tool to evaluate the employee's performance. As mentioned by one key informant*: "passing the quizzes can assess the employee behavior and level of awareness ….".* Lastly, by using the KPIs technique, it is possible to identify the number of training sessions/programmes the employees attended and completed. As a key informant explains: *"… we need to convince the management that the programme is doing great, and that employee behaviour is being changed. So, KPIs could be used to evaluate them".*
These tools, therefore, assist in the evaluation of employee performance with regard to SETA programmes and this also provides an indication of the programme's success.

In comparing these findings with those presented in the literature, it was noted that there are several studies which discuss the use of evaluations for the SETA programme. For example, Rantos et al (2012) illustrate several methods for evaluating a SETA programme. One of those methods is using a survey / questionnaire to evaluate the success of the programme overall. Other methods evaluate security awareness campaigns by highlighting that gaps exist and measuring the effectiveness of the SETA programme (Alshaikh et al., 2018; Johnson, 2006). However, this is an area that requires further research.

11. CSF#11: Measure Employee Reporting of Security Incidents

This CSF highlights the security incidents reported by the employee. Most organizations use phishing campaigns to simulate attacks. They want to know how many of the employees click the suspicious links, to measure the employees' awareness and knowledge regarding IS security issues. Thus, an increase in the number of suspicious links or other incidents reported by the employees is a valuable indication of the SETA programme's effectiveness. Within this research study, key informants described the methods to evaluate employee behaviour and the level of their awareness regarding the detection and reduction in security incidents. When the employee sends emails to the IS security department to report a suspicious link, that reflects on the success of the SETA programme. For example, one key informant comments: *"the reporting of a suspicious email indicated they get the awareness message*". The employees are the strongest link to protect the organization, provided they are aware of the suspicious emails and report them directly. In addition, the KPI tool can also be used to compare the current and previous years to measure the percentage of clicks on suspicious links. If employees recognize a percentage decrease in clicks, then it shows that the SETA programme is effective and improving security. As mentioned by one key informant: "*KPIs as a tool will let you know percentages and statistics, e.g., how many people clicked on suspicious links….".* Lastly, most organizations rely on phishing campaigns, as a key informant states: *"a simulation phishing campaign is used to identify who clicks and opens suspicious emails, and the percentage of those who report the incident to the security department…".* The main reason for a phishing simulation is to raise the level of awareness among employees. Therefore, reducing the number of security incidents (e.g. clicks on suspicious links) would show that the level of awareness is increasing (highlighting SETA programme effectiveness).

In comparing these findings with those presented in the literature, a number of observations can be made. Several studies recommend various countermeasures

that can be used to reduce IS security incidents (c.f. Chen et al., 2015; D'Arcy et al., 2009; Peltier, 2005). For example, D'Arcy et al., (2009) proposes that a SETA programme aims to mitigate IS risks and security incidents. Understanding the IS security policies through the delivery of SETA reduces IS security misuse (Peltier, 2005). It can be argued that a decreasing number of security incidents and security attacks provides an organization with a significant indication that the practice improvements are due to a successful SETA programme.

V. CONCLUSIONS AND FUTURE RESEARCH

This paper presents an exploratory study identifying the CSFs for SETA programmes. The CSFs emerge from the analysis of 20 key informant accounts of SETA programme effectiveness. The 11 CSFs are associated with the design, development, implementation, and evaluation phases of a SETA programme lifecycle. We found six CSFs relating to the **design** phase (CSF#1,2,3,4,5,6), one CSF relating to the **development** phase (CSF#7), two CSFs relating to the **implementation** phase (CSF#8,9), and two CSFs relating to the **evaluation** phase (CSF#10,11). The next step in this research is to conduct a focus group with additional key informants (experts) who have valuable experience in SETA programmes. The purpose of this next step is to validate our findings and to rank the 11 CSFs in order of importance. These findings will further contribute to building a lifecycle model of CSFs for SETA programmes.

VI. REFERENCES

1) Alina Ali Zani, A., Anir Norman, A., & Abdul Ghani, N. (2018). A Review of Security Awareness Approach: Ensuring Communal Learning. *PACIS 2018 Proceedings*, 278. Retrieved from https://aisel.aisnet.org/pacis2018/278
2) Alshaikh, M., Maynard, S. B., & Ahmad, A. (2021). Applying social marketing to evaluate current security education training and awareness programs in organisations. *Computers & Security*, *100*, 102090. https://doi.org/10.1016/j.cose.2020.102090
3) Alshaikh, M., Maynard, S. B., Ahmad, A., & Chang, S. (2018). An Exploratory Study of Current Information Security Training and Awareness Practices in Organizations. *Proceedings of the 51st Hawaii International Conference on System Sciences*, *9*, 5085 5094.https://doi.org/10.24251/hicss.2018.635
4) Alshaikh, M., Naseer, H., Ahmad, A., Maynard, S. B., paper Alshaikh, R., & Sean, M. (2019). Toward Sustainable Behaviour Change: An Approach for Cyber Security Education Training and Awareness. *Twenty-Seventh European Conference on Information Systems (ECIS2019),* 0–14. Retrieved from https://aisel.aisnet.org/ecis2019_rp/100
5) Alyami, A., Sammon, D., Neville, K., & Mahony, C. (2020). Exploring IS security themes: a literature analysis. *Journal of Decision Systems*, *29*(sup1), 425-437. https://doi.org/10.1080/12460125.2020.1848379
6) Barlow, J.B., Warkentin, M., Ormond, D., & Dennis, A.R. (2018). Don't even think about it! The effects of antineutralization, informational, and normative communication on information security compliance. Journal of the Association for Information Systems, 19(8), 689–715. https://doi.org/ 10.17705/1jais.00506
7) Bauer, S., Bernroider, E. W. N., & Chudzikowski, K. (2017). Prevention is better than cure! Designing information security awareness programs to overcome users' non-compliance with information security policies in banks. *Computers and Security*, *68*, 145–159. https://doi.org/10.1016/j.cose.2017.04.009
8) Bhattacherjee, A. (2012). Social science research: Principles, methods, and practices (2nd ed.). Tampa, FL: CreateSpace Independent Publishing Platform.
9) Bulgurcu, B., Cavusoglu, H., & Benbasat, I. (2010). Information security policy compliance: An empirical study of rationality-based beliefs and information security awareness. MIS Quarterly, 34(3), 523–548. https://doi.org/10.2307/25750690
10) Bullen, C. V., and Rockart, J. F. (1981). A primer on critical success factors.
11) Burns, A. J., Roberts, T. L., Posey, C., Bennett, R. J., and Courtney, J. F. (2018). Intentions to Comply Versus Intentions to Protect: A VIE Theory Approach to Understanding the Influence of Insiders 'Awareness of Organizational SETA Efforts. *Decision Sciences*, *49*(6), 1187–1228. https://doi.org/10.1111/deci.12304contributions to outcomes research", Circulation, Vol. 119 No. 10, pp. 1442-1452.

12) Chen, Y. A. N., Ramamurthy, K. R. A. M., & Wen, K. W. (2015). Impacts of comprehensive information security programs on information security culture. *Journal of Computer Information Systems*, *55*(3), 11-19. http:10.1080/08874417.2015.11645767
13) Cram, W.A., D'Arcy, J., and Proudfoot, J.G. (2019). Seeing the forest and the trees: A meta-analysis of the antecedents to information security policy compliance. MIS Quarterly, 43(2), 525–554. https://doi.org/10.25300/MISQ/2019/15117
14) D'Arcy, J., Herath, T., Yim, M.-S., Nam, K., & Rao, H. R. (2018). Employee Moral Disengagement in Response to Stressful Information Security Requirements: A Methodological Replication of a Coping-Based Model. *AIS Transactions on Replication Research*, *4*(June), 1–18. https://doi.org/10.17705/1atrr.00028
15) D'Arcy, J., Hovav, A., & Galletta, D. (2009). User awareness of security countermeasures and its impact on information systems misuse: A deterrence approach. *Information Systems Research*, *20*(1), 79 98.https://doi.org/10.1287/isre.1070.0160
16) De Maeyer, D. (2007). Setting up an effective information security awareness programme. *ISSE/SECURE 2007 - Securing Electronic Business Processes: Highlights of the Information Security Solutions Europe/SECURE 2007 Conference*, (2007), 49–58. https://doi.org/10.1007/978-3-8348-9418-2_5
17) Dhillon, G., Talib, Y. Y. A., & Picoto, W. N. (2020). The mediating role of psychological empowerment in information security compliance intentions. *Journal of the Association for Information Systems*, *21*(1), 152–174. https://doi.org/10.17705/1jais.00595
18) Han, J. Y., Kim, Y. J., & Kim, H. (2017). An integrative model of information security policy compliance with psychological contract: Examining a bilateral perspective. Computers and Security, 66, 52–65. https://doi.org/10.1016/j.cose.2016.12.016
19) Hansche, S. (2001). Designing a security awareness program: Part 1. *Information systems security*, *9*(6), 1-9. http:10.1201/1086/43298.9.6.20010102/30985.4
20) Heikka, J. (2008). A constructive approach to information systems security training: An action research experience. *14th Americas Conference on Information Systems, AMCIS 2008*,*1*,15–22. https: https://aisel.aisnet.org/amcis2008/319/
21) Herath, T., & Rao, H. R. (2009). Encouraging information security behaviors in organizations: Role of penalties, pressures and perceived effectiveness. *Decision Support Systems*, *47*(2), 154–165. https://doi.org/10.1016/j.dss.2009.02.005
22) Herath, T., Yim, M. S., D'Arcy, J., Nam, K., & Rao, H. R. (2018). Examining employee security violations: moral disengagement and its environmental influences. *Information Technology and People*, *31*(6), 1135–1162. https://doi.org/10.1108/ITP-10-2017-0322
23) Hovav, A., & D'Arcy, J. (2012). Applying an extended model of deterrence across cultures: An investigation of information systems misuse in the US and South Korea. *Information & Management*, *49*(2), 99110.http://europepmc.org/abstract/med/10297607
24) Siqi Hu, Carol Hsu & Zhongyun Zhou (2021a) Security Education, Training, and Awareness Programs: Literature Review, Journal of Computer Information Systems, DOI: 10.1080/08874417.2021.1913671
25) Hu, S., Hsu, C., & Zhou, Z. (2021b). The impact of SETA event attributes on employees' security-related Intentions: An event system theory perspective. *Computers & Security*, *109*, 102404. https://doi.org/10.1016/j.cose.2021.102404
26) Johnson, E. C. (2006). Security awareness: Switch to a better programme. *Network Security*,*2006*(2),15–18. https://doi.org/10.1016/S1353-4858(06)70337-3
27) Karjalainen, M. and Siponen, M.T. (2011), "Toward a new meta-theory for designing information systems (IS) security", Journal of the Association for Information Systems, Vol. 12 No. 8, pp. 518-555. Retrieved from http://sprouts.aisnet.org/9-53 systems (IS) security", Journal of the Association for Information Systems, Vol. 12 No. 8, pp. 518-555.
28) Karjalainen, M., Siponen, M., Puhakainen, P., & Sarker, S. (2013). One Size Does Not Fit All: Different Cultures Require Different Information Systems Security Interventions. *PACIS 2013 Proceedings*, Paper 98.

29) Kawulich, B. B. (2004). Data analysis techniques in qualitative research. *Journal of research in education*, *14*(1), 96-113.
30) Kirova, D., & Baumoel, U. (2018). Factors that Affect the Success of Security Education, Training, and Awareness Programs: A Literature Review. *Journal of Information Technology Theory and Application (JITTA)*, *19*(4), 4.
31) Lebek, B., Uffen, J., Neumann, M., Hohler, B., & Breitner, M. H. (2014). Information security awareness and behavior: A theory-based literature review. *Management Research Review*, *37*(12), 1049–1092. https://doi.org/10.1108/MRR-04-2013-0085
32) Leech, N. L., and Onwuegbuzie, A. J. (2007). An array of qualitative data analysis tools: A call for data analysis triangulation. *School psychology quarterly*, *22*(4), 557.
33) Mahmood, M.A., Siponen, M., Straub, D., Rao, H.R., & Raghu, T.S. (2010). Moving toward black hat research in information systems security: An editorial introduction to the special issue. MIS Quarterly, 34(3), 431–433. https://doi.org/10.2307/25750685
34) Marshall, C., and Rossman, G. (1989). Designing Qualitative Research. Newbury Park, CA: Sage Publications
35) Marshall, M. N. (1996). The key informant technique. *Family Practice*, *13*(1), 92–97. https://doi.org/10.1093/fampra/13.1.92
36) Okenyi, P. O., and Owens, T. J. (2007). On the anatomy of human hacking. *Information Systems Security*, *16*(6),302-314. https:10.1080/10658980701747237
37) Parrish, J. L., & San Nicolas-Rocca, T. (2012). Toward Better Decisions with Respect to Is Security: Integrating Mindfulness Into IS Security Training. *Pre-ICIS Workshop on Information Security and Privacy (SIGSEC)*, 1–16. Retrieved from http://aisel.aisnet.org/wisp2012/17
38) Pastor, V., Díaz, G., & Castro, M. (2010). State-of-the-art simulation systems for information security education, training and awareness. *2010 IEEE Education Engineering Conference, EDUCON 2010*, 1907–1916. https://doi.org/10.1109/EDUCON.2010.5492435
39) Patton, M. Q. (1990). *Qualitative evaluation and research methods*. SAGE Publications, inc.
40) Peltier, T. R. (2005). Implementing an information security awareness program. *Information Systems Security*, *14*(2), 37–49. https://doi.org/10.1201/1086/45241.14.2.20050501/88292.6pp. 1758-1772 Publications, Newbury Park, CA.
41) Puhakainen, P., & Siponen, M. (2010). Improving employees' compliance through information systems security training: An action research study. MIS Quarterly, 34(4), 757–778. https://doi. org/10.2307/25750704
42) Rantos, K., Fysarakis, K., & Manifavas, C. (2012). How effective is your security awareness program? An evaluation methodology. *Information Security Journal: A Global Perspective*, *21*(6), 328-345. research: developing taxonomy, themes, and Retrieved
43) Rogers, R. (1975). A Protection Motivation Theory Of Fear Appeals And Attitude Change. *Journal of Psychology: Interdisciplinary and Applied*. https://doi.org/10.1080/00223980.1975.9915803
44) Silic, M., & Lowry, P. B. (2020). Using design-science based gamification to improve organizational security training and compliance. *Journal of Management Information Systems*, *37*(1), 129-161. https://doi.org/10.1080/07421222.2019.1705512
45) Siponen, M. T. (2000). A conceptual foundation for organizational information security awareness. Information Management & Computer Security, 8(1), 31–41. https://doi.org/10.1108/09685220010371394
46) Siponen, M., & Vance, A. (2010). Neutralization: New insights into the problem of employee information systems security policy violations. MIS Quarterly, 34(3), 487–502. https://doi.org/10.2307/25750688
47) Straub, D. W., & Welke, R. J. (1998). Coping with systems risk: Security planning models for management decision making. *MIS Quarterly: Management Information Systems*, *22*(4), 441–464. https://doi.org/10.2307/249551
48) Strauss, A., & Corbin, J. (1990). Basics of qualitative research: Grounded theory procedures and techniques. Thousand Oaks, CA: SAGE Publications, Inc. theory", Health Services Research, Vol. 42
49) Tsohou, A., Karyda, M., Kokolakis, S., & Kiountouzis, E. (2015). Managing the

introduction of information security awareness programmes in organisations. *European Journal of Information Systems*, *24*(1), 38–58. https://doi.org/10.1057/ejis.2013.27

50) Von Solms, R., & Von Solms, B. (2004). From policies to culture. *Computers & security*, *23*(4), 275-279 https://doi.org/10.1016/j.cose.2004.01.013

51) Vroom, C., & Solms, R. V. (2002). A practical approach to information security awareness in the organization. In *Security in the Information Society* (pp. 19-37). Springer, Boston, MA. https: 10.1007/978-0-387-35586-3_46
52) Walsham, G. (2002). Cross-cultural software production and use: a structurational analysis. *MIS quarterly*, 359-380.
53) Walsham, G. (2006). Doing Interpretive Research. European Journal of Information Systems, 15(3),320–330. Retrieved from https://link.springer.com/article/10.1057/palgrave.ejis.3000589
54) Whitman, M.E., and Mattord, H.J. 2008. Principles of Information Security. Stamford, Connecticut: Course Technology.
55) Wilson, M., & Hash, J. (2003). Building an information technology security awareness and training program. *NIST Special publication*, *800*(50), 1-39.
56) Wu He & Zuopeng (Justin) Zhang (2019) Enterprise cybersecurity training and awareness programs: Recommendations for success, Journal of Organizational Computing and Electronic Commerce, 29:4, 249-257, DOI: 10.1080/10919392.2019.1611528
57) Yoo, C. W., Sanders, G. L., & Cerveny, R. P. (2018). Exploring the influence of flow and psychological ownership on security education, training and awareness effectiveness and security compliance. *Decision Support Systems*, *108*(February), 107–118. https://doi.org/10.1016/j.dss.2018.02.009

Improving Resistance of Matrix Factorisation Recommenders To Data Poisoning Attacks

Sulthana Shams
School of Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland
sshams@tcd.ie

Douglas J. Leith
School of Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland
doug.leith@tcd.ie

***Abstract*—In this work, we conduct a systematic study on data poisoning attacks to Matrix Factorisation (MF) based Recommender Systems (RS) where a determined attacker injects fake users with false user-item feedback, with an objective to promote a target item by increasing its rating. We explore the capability of a MF based approach to reduce the impact of attack on targeted item in the system. We develop and evaluate multiple techniques to update the user and item feature matrices when incorporating new ratings. We also study the effectiveness of attack under increasing filler items and choice of target item.**

Our experimental results based on two real-world datasets show that the observations from the study could be used to design a more robust MF based RS.

***Index Terms*—recommender systems, matrix factorisation, data poisoning attacks, attack resistance**

I. Introduction

The issue of robustness against malicious attack is receiving attention from the research community [1, 2, 3]. Recommender System (RS) in social media platforms such as Facebook and Twitter have been in the limelight due to the risks they constantly pose to society by influencing their user base. From the point of view of RS, not only do they recommend items by learning a user's preferences but also help users discover and develop new interests thus influencing user behavior.

In poisoning attacks, an adversary creates fake profiles with carefully crafted ratings for items and attempts to target an item with the objective of increasing or decreasing the item's rating, thus making the item more/less likely to be recommended by the system. For our work, we consider that the attacker's goal is to promote a target item, i.e. an attacker-chosen target item's rating is increased and thus is more likely to be recommended to true users. We look at a common attack strategy called 'Average Attack' on collaborative filtering systems discussed in literature [1, 4, 5]. We assume that the attacker can only inject a limited number of fake users and each fake user rates a limited number of items (including the target item and other non-target items called filler items) to evade suspicion.

In this paper, we revisit Matrix Factorisation (MF) based RS [6]. MF is widely known in RS due to its simplicity and effectiveness. The typical paradigm of MF in RS is to decompose the user-item interaction matrix $R \in \mathbb{R}^{m \times n}$ into the product of two low-dimensional latent matrices $U \in \mathbb{R}^{d \times m}$ and $V \in \mathbb{V}^{d \times n}$ such that their dot product $U^T \cdot V$ is a good approximation of R. Matrix U captures the relationship between a user and the latent features while matrix V captures the relationship between an item and the features. We call U as the user-feature matrix and V as the item-feature matrix.

Typically, when new ratings are introduced to the system, the latent feature matrices are updated to incorporate the new ratings and thus update the user-item prediction matrix. Most works in literature take random items or unpopular items with fewer ratings as target items [3, 7, 8, 9]. We consider target items with different number of ratings received by true users and look at the shift in rating of the target item after updates to U, V. We conclude that some items are easier to attack than others. Items with fewer ratings are most vulnerable to attacks presumably due to the ease with which their feature vectors in V can be changed. An item which received a large number of ratings from the true users proves harder to attack.

Based on these observations, we further explore the role of U and V as a possible defense mechanism against fake user attacks. While one common approach is regular MF where both U and V are adapted, we look at other ways to boost the recommendation robustness under data poisoning attacks by looking at different ways of updating these latent feature matrices when introduced to new ratings. For example, consider the following new ways to incorporate the newly added ratings by fake users :

i) Hold V constant and update U just for the fake users. Here attack has no effect on true users. The U for the true users remains same as before the attack. Although this offers an immunity to attacks, it has no collaborative learning involved since U and V of true users remain unchanged to any incoming ratings.

ii) Hold V constant, add attackers and update U for all users.

iii) Thirdly, perform a modified alternating least squares method where find U using (i), then update V and repeat until converged.

From our study, we observe that ii) leads to very low change in rating of target item after attack. In comparison iii) shows larger change in rating of target item. So the effect of the

This work was supported by SFI grant 16/IA/4610

attack is pronounced when V is updated.

We show that these observations could be used to make updates to predicted ratings matrix more robust.

II. Related Work

The impact of data poisoning attacks where fake users are injected in RS with carefully crafted user-item interaction has been studied extensively. Detailed survey on attack models and robustness of RS algorithms are provided in [1, 2, 4, 5].

Recently, there is a line of work [7, 8, 9, 10] focusing on modelling the attack as an optimisation problem to decide the rating scores for the fake users and model attacks specific to the type of RS. For example, [8, 9] proposes data poisoning attacks for deep learning based RS and graph-based RS respectively. [7] proposes to select a subset of true users who are influential to the recommendations, to craft ratings for the fake user's attack on regular MF based RS.

In [10], instead of attacking the top-N recommendation lists, their goal was to study the change in the rating predictions after attack, for all missing entries of the rating matrix.

Most works in literature [7, 8, 9] use HR@N or 'Hit -Ratio' as the metric to study the effectiveness of attack where Hit-Ratio of a target item is the fraction of normal users whose top-N recommendation lists contain the target item.

We feel that the top-N recommendation list per user is too fragile a metric for observing attack effectiveness since the relevance of that list is user dependent. The standard prediction shift metric [1, 4, 11] used in literature also seem crude. It does not account for the initial rating of the target item before attack. i.e a target item with low initial rating would show larger deviation than an item with rating closer to mean value before attack. Keeping in mind all of the above, we introduce a new metric that gives the change in rating relative to the maximum deviation possible after attack. i.e. It depicts the ratio of the maximum deviation that the attack has achieved.

Although the impact of filler items in attack effectiveness in terms of Hit Ratio is studied in [8, 9], no relationship between the hit ratio and the number of filler items was concluded. The relationship was shown to be heavily dependent on the data-sets. Interestingly, our study on the same using the relative change in mean metric yielded a different result. Increasing filler items also increased the relative change in rating of the target item under attack.

While there are many studies exploring defensive techniques against data poisoning attacks [12, 13, 14], to the best of our knowledge, there is no existing studies that look at the factors affecting the defence capability of MF based RS.

III. Attack Model

For the type of attacks that we focus on, there is a *target item* that the attacker is interested in promoting and a set of *filler items* that is used to make the fake users seem real and ensure that some correlation is established with other true users.

A. Target Item

The target item is given the maximum rating to promote it in the system. We consider three types of target items based on the number of ratings received from the true users. Specifically, in our experiments, we sample an item uniformly at random from those items which have received 1, 10, 100 ratings and treat it as the target item.

B. Attack Knowledge

We assume that the adversary knows the mean rating and standard deviation for every item in the system. This is a reasonable assumption since such aggregate information about user preferences may be found online from databases which publicly displays the average user ratings of items. (e.g. movie databases, amazon product databases etc)

The filler items are chosen randomly from the list of items. Intuitively it will be much more difficult to detect such a fake user profile since the set of rated items change from profile to profile. The ratings for the filler items are sampled from the Gaussian distribution using the mean rating and standard deviation of every item available with the adversary.

IV. Experiments

A. Datasets

We evaluate the effectiveness of attack on the MovieLens dataset (943 users rating 1682 movies, contains 100000 ratings from 1-5) which is widely used in literature for evaluating recommender systems under attack and Goodreads 10K dataset (53,424 people rating 10,000 books, 5.9M ratings from 1-5).

We take a dense subset of the Goodreads dataset, obtained by selecting the top 1000 users which have provided the most ratings. This provides us with 1000 users and 8557 items rated by these top 1000 users.

B. Evaluation Setup

Unless mentioned otherwise, the attack size is fixed to 1% of the total true user population. We also look at how the number of filler items and the number of ratings of targeted item by true users impact attack effectiveness. We sample 50 instances of target items under each set-up and will average their experimental results.

C. Performance Metrics

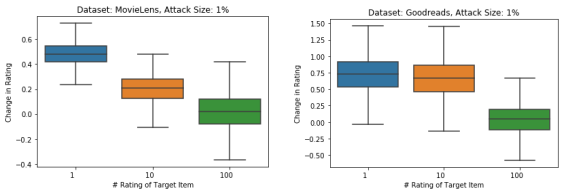
We use change in rating of target item relative to the maximum deviation possible as our evaluation metric.

$$\text{Change in Rating}_u = \frac{\mu_f(u,i) - \mu_o(u,i)}{|5 - \mu_o(u,i)|}$$

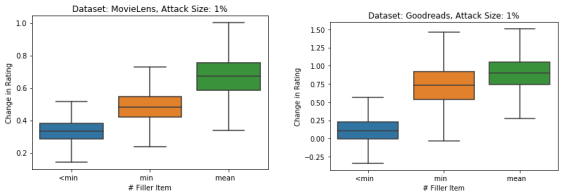
where $\mu_f(u,i)$ is the predicted rating of target item i of user u after attack, $\mu_o(u,i)$ is rating of the same target item i of user u before attack and 5 is the maximum rating that can be given to target item.

V. Results Analysis

Let us first consider the usual MF where we adapt both U and V and look at how number of attacker filler items and ratings of target item by true users impact the attack.

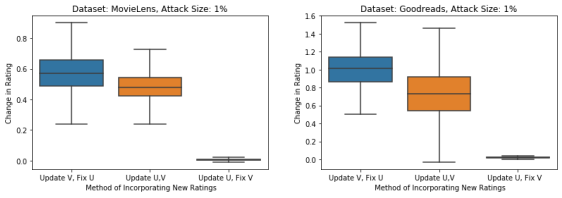


(a) (b)



(c) (d)

Fig. 1: Box-plot comparing the distribution of change in rating over true users for number of ratings of target item and different number of filler items respectively for MovieLens and Goodreads dataset when updating both U and V



(a) (b)

Fig. 2: Boxplot comparing the distribution of change in rating over true users for different methods of adapting new ratings using U, V for MovieLens and Goodreads dataset.

1) Impact of the number of ratings of target item : Figure 1 (a),(b) have box-plots showing distribution of change in rating when targeting items with different number of ratings. For this study, we fix the number of filler items to the minimum number of items rated by true users in the respective data-set.

It can be seen from both Figure 1(a) and Figure 1(b) that some items are easier to attack than others. From table I, an item with just 1 rating shows the most mean shift in rating after attack compared to an item with 100 ratings when both U, V are updated. In fact, target item with 100 ratings proves much harder to attack with a mean change in rating close to 0 for both data-sets.

Targeting items with higher number of ratings seems to make it difficult for the attackers to change the feature vector. In comparison, an item with fewer ratings is fragile and an attack to such an item would be very difficult to defend. It becomes extremely easy for the attackers to change the feature vector of such an item.

2) Impact of the number of filler items: Figure 1 (c),(d) show the impact of the number of filler items on our attacks for target item.

	# target ratings		# filler items	
Dataset/Update Method	1	100	<min	=mean
ML/ Update U,V	0.48	-0.04	0.33	0.67
ML/ Fix V, Update U	0.005	0.04	0.005	0.005
GR/ Update U,V	0.73	-0.005	0.10	0.89
GR/ Fix V, Update U	0.02	0.1	0.02	0.02

TABLE I: Mean change in rating for MovieLens and Goodreads datasets for number of ratings of target item =1,100 and number of filler items=minimum, mean number of items rated by true users. Legend: ML=MovieLens dataset, GR=Goodreads dataset

For this study we fix the item with 1 rating as the target item for both MovieLens and Goodreads data-set since it is the easiest to attack.

For both data-set, we observe the distribution of change in rating against number of filler items as 1) less than the minimum number of items rated by true users in the data-set 2) equal to minimum number of items rated by true users in the data-set 3) equal to the mean number of items rated by true users in the data-set.

It can be seen from both Figures 1 (c) and (d) and table I that increasing filler items from less than minimum to mean increases the mean change in rating of target item when both U, V are updated. It seems that changes to V are more pronounced when filler items increase. Perhaps because such a fake user would be similar to more true users and thus contributes more to change in V.

Although an attacker would achieve increased change in rating of target item when using more filler items, rating items more than the mean number of items rated by true users may be flagged as a suspicious behaviour.

For the rest of the experiment, we fix the number of filler items to the minimum number of items rated by true users in the data-set and choose a target item with one rating to better capture the effects of attack for the next part of the experiment.

A. Incorporating New Ratings

Figure 2 (a),(b) compares the distribution of change in rating over true users in the data-set when applying different ways to adapt U and V vectors to incorporate new ratings into the system. As discussed previously, for all the scenarios below we fix the number of filler items to the minimum number of items rated by true users in the data-set and choose a target item with one rating to show our results.

1) Update V and fix U: In this set-up, we first update U only for attackers by holding V constant. We obtain an updated U for fake users but with values for true users same as before. Then proceed to update V.

Here, change in rating observed after attack is slightly higher for both the data-sets in comparison to regular MF. The effect of number of ratings of target item and number of filler items are similar to Figures 1 (a),(b) and so are not reported separately. Just as in regular MF, a fragile item with one rating can be easily attacked while a well-reviewed item proves harder to attack.

2) Update U and fix V: We hold V constant, add attackers and update U for all users. From Figure 2, attack has only a small effect on ratings for true users in this scenario. Updated U is very close to its original version before attack assuming that the regularisation penalty is not too high. This means the change in rating on target item after attack is very low.

The effect of number of filler items and ratings of target items is reported in table I. The increasing number of filler items seems to have no effect on this set-up. Also, the mean change in rating for any choice of target item are found to be negligible compared to regular MF. We believe that the small increase in change in rating as we move from target item with 1 rating to 100 ratings for this set-up comes from the regularisation penalty applied.

We conclude that the effect of the attack on the target item is captured by V matrix. As long as V is kept constant, updated U after attack is very similar to the one before attack unless regularisation parameter is too high. This ensures that the predicted ratings matrix after attack is very close to the predicted matrix before attack. So effect of attack on true users is found negligible.

Thus using method 2) for incremental updates to U when new ratings are added, then periodically using regular MF to update U and V would help in monitoring attacks. If a big difference between their results is observed, then that might flag a warning for items that change rating a lot.

B. Conclusions

In this paper, we revisited the MF approach to RS and studied the effect of attack under different update methods of latent matrices when incorporating new ratings. We also studied the effectiveness of attack under increasing filler items and choice of target item.

We can use these observations to make updates to RS more robust. Items that are more vulnerable to attacks can perhaps be defended from fake users by using dummy ratings which would make it harder for injected fake users to change their feature vector. Also updates to latent feature matrices need not be performed frequently together. Instead, regular MF methodology could be used periodically with Approach 2 for incremental updates to the ratings matrix. Thus any large shift in rating of items could be monitored periodically and necessary actions taken.

Our approaches are simple, yet effective and can be easily used in existing systems.

REFERENCES

[1] S. K. Lam and J. Riedl, "Shilling recommender systems for fun and profit," in *Proceedings of the 13th International Conference on World Wide Web*, ser. WWW '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 393–402. [Online]. Available: https://doi-org.elib.tcd.ie/10.1145/988672.988726

[2] B. Mobasher, R. Burke, R. Bhaumik, and C. Williams, "Toward trustworthy recommender systems," *ACM Transactions on Internet Technology*, vol. 7, pp. 23–es, 10 2007.

[3] C. Wu, D. Lian, Y. Ge, Z. Zhu, E. Chen, and S. Yuan, *Fight Fire with Fire: Towards Robust Recommender Systems via Adversarial Poisoning Training*. New York, NY, USA: Association for Computing Machinery, 2021, p. 1074–1083. [Online]. Available: https://doi-org.elib.tcd.ie/10.1145/3404835.3462914

[4] K. Patel, AmitThakkar, C. Shah, and K. Makvana, "A state of art survey on shilling attack in collaborative filtering based recommendation system," 11 2015.

[5] S. Mingdan and Q. Li, "Shilling attacks against collaborative recommender systems: a review," *Artificial Intelligence Review*, vol. 53, 01 2020.

[6] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[7] M. Fang, N. Z. Gong, and J. Liu, *Influence Function Based Data Poisoning Attacks to Top-N Recommender Systems*. New York, NY, USA: Association for Computing Machinery, 2020, p. 3019–3025. [Online]. Available: https://doi-org.elib.tcd.ie/10.1145/3366423.3380072

[8] H. Huang, J. Mu, N. Z. Gong, Q. Li, B. Liu, and M. Xu, "Data poisoning attacks to deep learning based recommender systems," *Proceedings 2021 Network and Distributed System Security Symposium*, 2021. [Online]. Available: http://dx.doi.org/10.14722/ndss.2021.24525

[9] M. Fang, G. Yang, N. Z. Gong, and J. Liu, "Poisoning attacks to graph-based recommender systems," *Proceedings of the 34th Annual Computer Security Applications Conference*, Dec 2018. [Online]. Available: http://dx.doi.org/10.1145/3274694.3274706

[10] B. Li, Y. Wang, A. Singh, and Y. Vorobeychik, "Data poisoning attacks on factorization-based collaborative filtering," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 1893–1901.

[11] B. Mobasher, R. Burke, R. Bhaumik, and J. Sandvig, "Attacks and remedies in collaborative recommendation," *IEEE Intelligent Systems*, vol. 22, no. 3, pp. 56–63, 2007.

[12] C. Williams, B. Mobasher, and R. Burke, "Defending recommender systems: Detection of profile injection attacks," *Service Oriented Computing and Applications*, vol. 1, pp. 157–170, 10 2007.

[13] P.-A. Chirita, W. Nejdl, and C. Zamfir, "Preventing shilling attacks in online recommender systems," in *Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management*, ser. WIDM '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 67–74. [Online]. Available: https://doi-org.elib.tcd.ie/10.1145/1097047.1097061

[14] R. Bhaumik, C. Williams, B. Mobasher, and R. Burke, "Securing collaborative filtering against malicious attacks through anomaly detection," 01 2006.

Insecure Software on a Fragmenting Internet

Ita Ryan
School of Computer Science
University College Cork
Cork, Ireland
ita.ryan@cs.ucc.ie

Utz Roedig
School of Computer Science
University College Cork
Cork, Ireland
utz.roedig@cs.ucc.ie

Klaas-Jan Stol
School of Computer Science
University College Cork, Lero
Cork, Ireland
k.stol@cs.ucc.ie

***Abstract*—Global geopolitical forces are pushing much of the world towards Internet nationalism, threatening to turn the Internet into a 'Splinternet.' In this paper we argue that the crisis in software security will exacerbate this trend. We examine existing moves towards Internet fragmentation on multiple levels. We discuss current trends in online crime, espionage, and warfare. We look at the role of software vulnerabilities, discussing how the prevalence of software security issues could propel nations further apart. We argue that there is an urgent need for a 'zero tolerance' attitude to software security issues, and discuss what is needed to create this.**

***Index Terms*—Cybersecurity, software security, Internet nationalism, Splinternet**

I. Introduction

With its elegant protocols and built-in redundancy, the Internet is inherently global in nature. Nevertheless, it is not immune to geopolitical forces. Inter-country fragmentation is happening on several different levels, and has been referred to as the 'Splinternet' [1].

Ubiquitous access to the Internet means that software flaws can be exploited remotely from anywhere, with local law enforcement having no jurisdiction in the country from which a crime was committed. Thus, the Internet facilitates previously unimaginable scenarios like the May 2021 ransomware attack on the Irish Health Service Executive [2]. Similarly, espionage, sabotage, and cyberwarfare can be conducted remotely, providing hostile forces with unprecedented access.

Secure software is a core cybersecurity concern. While firewalls, anti-virus tools, network segmentation, and other tools and strategies are deployed to protect digital assets, software defects and design flaws can provide attackers with a back door. It is impossible to prove the security of non-trivial software. Indeed, severe implementation flaws have been found in firewalls [3], anti-virus tools [4] and network segmentation tools [5] themselves.

The number of newly reported software vulnerabilities increases each year [6]. Efforts to tackle software security issues are haphazard. Until very recently there was little government guidance, and organisational software security drives in unregulated industries are entirely voluntary. While critical domains use regulations often based on the U.S. National Institute of Standards and Technology (NIST) guidelines, these guidelines are rather heavy-weight, and thus unsuitable for most organisations. We argue that a rapid escalation of effort in eliminating software vulnerabilities is needed. Otherwise, exploitation will continue to increase, exacerbating the trend towards Internet nationalism. The vulnerability of military and critical infrastructure and of nuclear control software to remote exploitation will be seen as too much of a risk.

Previously, Claessen [7] discussed how the understanding of cyberspace as a military as well as civilian domain has led to increasing attempts to impose state sovereignty on the Internet, with particular reference to the different approaches adopted by Russia and the European Union (EU). Hoffman looked at how the new technical standards proposed by China could lead to Internet fragmentation [1]. In this paper, we contribute to this line of work by examining contemporary pressures on a cohesive Internet, explore the forces that are driving the Internet to fragment, and consider how untamed software security risk adds to those pressures. We advocate for a new culture of software insecurity intolerance.

In Section II we look at drivers towards Internet nationalism and ways in which countries are currently uncoupling from a cohesive global Internet. In Section III we discuss large scale security issues that the Internet facilitates. In Section IV we examine how software vulnerabilities impact on cybersecurity. In Section V we discuss approaches to reducing software vulnerabilities, and some exacerbating factors. The conclusion in Section VI discusses possible global consequences of a failure to improve software security.

II. Internet Nationalism

Internet fragmentation is an existing phenomenon driven by perceived national interest and facilitated by design choices on different Internet layers. We first discuss the different layers in which changes are happening. We then briefly discuss how some countries are diverging on multiple levels.

A. OSI Model Layer 1: Physical

Approximately 95% of global Internet traffic travels through undersea fibre-optic cables, which comprise the Internet's backbone [8]. Cables are increasingly perceived as relevant to geopolitical tensions [9]. Russian naval exercises off the Irish coast in January 2022 focused minds on the vulnerability of transatlantic communications cables, damage to which would severely impair Irish and European Internet connectivity [10]. Underlining the fragility of the world's Internet connectivity, in January 2022 Tonga's external communications were almost

completely cut off after a volcanic explosion severed the single undersea cable connecting it to Fiji [8].

Russia recently decreed that its transnational cables must be registered with a central authority [11]. Data on transnational cables is already collated and made public in the U.S. [12]. U.S. researchers recently mapped crucial internal cables in a project funded by the Dept. of Homeland Security [13].

There is concern about espionage via physical cable access [14], with China-funded cables increasingly regarded with suspicion [15]. The U.S. Department of Justice (DoJ), in 2020, objected on national security grounds to a new undersea cable connecting the U.S. to Hong Kong [16].

B. OSI Model Layer 2: Data Link

The U.S. have banned use of Chinese company Huawei's technology in 5G networks, citing security concerns [17]. Four other Chinese tech companies have also been deemed security threats by the U.S. Federal Communications Commission (FCC) [18]. Russia mandates use of local technology for key Internet controls [11]. As each country moves to using only local suppliers, commonality declines and the feasibility of standards and protocols diverging increases.

C. OSI Model Layer 3: Network

Communication between networks is often done via Internet Exchange Points (IXPs) where multiple network endpoints are located in close proximity, using Border Gateway Protocol (BGP) records to move traffic directly between networks. This may be done for example to avoid transit fees [19]. BGP can be used to prevent a nation's Internet traffic from travelling through another nation's territory. Russia has directed that Internet traffic should only be directed through approved IXPs registered with Roskomnadzor [7]. This policy is likely to keep Russian Internet traffic within the country.

Because BGP is the protocol that allows networks to find destinations, it can also be used for censorship. Pakistan accidentally propagated an incorrect YouTube destination to the global Internet when it banned YouTube in 2008 [20]. Ververis et al. [21] found that BGP configuration is one of the most widely-used tools for Internet censorship. Limonier et al. found that, over the six years prior to 2020, Internet traffic from disputed Donbas in Ukraine shifted to being routed almost entirely through Russia [22]. They concluded that routing can reflect geopolitical concerns.

D. OSI Model Layer 4: Transport

All Internet traffic currently uses TCP/IP. While the transition from IPV4 to IPV6 brings its own fragmentation concerns [23], China has proposed a 'decentralised Internet' model and associated entirely new protocol named New IP. It argues that the 50-year-old IP protocol is creaking under today's massive Internet use and new communication needs for technologies like virtual and augmented reality [24]. New IP facilitates centralised surveillance and control of the Internet, and is seen by some as entailing the loss of individual freedom to the state [1]. It is suggested that the protocol will not be adopted by the U.S. or its allies and that this could lead to a fragmentation into at least two separate versions of the Internet, with different countries or blocs using their own protocols. Hoffman et al. [1] note that, although involvement in Internet protocol standards committees is resource-intensive and expensive, nations should participate in order to ensure that their values are reflected.

E. Data, Applications and Access

China, Russia, and other states require data pertaining to their citizens to be stored within their borders [25]. The EU only allows data to be held overseas if certain privacy and protection guarantees are followed. Data localisation allows states to ensure that their data remains within their jurisdiction, but it also contributes to fragmentation.

Many countries have banned or restricted other countries' websites and applications for reasons of censorship, privacy or national security. For example, China's 'Great Firewall' prevents the use of Twitter, Facebook, Google, Signal and numerous other applications [26]. In 2020, India banned over 200 Chinese apps including Baidu, WeChat and Alipay, citing national security and surveillance concerns amid escalating border tensions [27]. Russia's February 2022 invasion of Ukraine was swiftly followed by a ban on Instagram, Facebook and other sites due to 'extremist activities.' In March 2022, the FCC added Russian anti-virus organisation Kaspersky, already banned from U.S. government networks, to its list of firms posing a security threat [28].

Governments may use strategies to control the flow of information over the Internet [29], often to limit foreign content. In a 2020 global, longitudinal study of Internet censorship, Niaki et al. [30] found the most censorship overall in Iran, South Korea, Saudi Arabia, Kenya, and India. India is the world's largest democracy, a reminder that censorship is not the sole preserve of authoritarian regimes. Conservative countries may resist open access to pornographic or gambling sites, seeing these as conflicting with national values.

Removal of Internet access is a favourite tool of oppressive regimes in times of turmoil. For example, most Internet access was lost for three days during the 2016 general election in Uganda [31]. In Belarus, where all Internet access is government controlled, there was a 61-hour Internet blackout during protests against a disputed Presidential election result in August 2020 [32]. In Myanmar in February 2021, new cybersecurity laws were introduced allowing mass censorship and surveillance after a military coup. In January 2022, Kazakhstan was subject to an Internet blackout amid anti-government protests about fuel charge hikes [33]. Those are a few examples among many; the Access Now activist group estimated that there were at least 155 Internet shutdowns in 29 countries in 2020 alone (https://www.accessnow.org/).

F. Multi-Level Divergence

China's Internet is a model for all nations, like Russia, that want to be able to disconnect from the global Internet at will. It has no foreign telecommunications companies within its borders. External connections are made via cables that pass

TABLE I
ENISA CYBER CRIME ACTORS & MOST COMMON ACTIVITIES 2020-2021

Level	Description
State-sponsored actors	Malware Espionage Supply chain compromise Disinformation\misinformation Cybercrime for monetary gain Sabotage (targeting of Industrial Control System (ICS)s) Cyber arms race
Cybercriminals	Ransomware Cryptojacking Malware Cybercrime-as-a-service DDoS, Web Attacks
Hacker-for-hire actors	Access-as-a-Service
Hacktivists	DDoS Sensitive data release Account takeovers

through the 'Great Firewall,' leave China, and connect with external IXPs on foreign soil [34].

Having banned most U.S. apps, China has very successful social media apps of its own. Chinese government organisations were ordered to remove foreign hardware and software from their offices by the end of 2022 in a 2019 edict [35]. This move away from reliance on computers and software developed by the U.S. and its allies, along with the Intranet-like nature of China's Internet, its use of the 'Great Firewall' and its drive to replace IP with New IP show that China is splitting from the global Internet on multiple levels.

Russia has been attempting to emulate China and modify its Internet (the 'RuNet') to remove dependence on external connections at every level. For example, in 2017 the Russian Security Council launched a process for developing a parallel DNS service [36]. A 2019 law mandated installation of local apps on devices sold in Russia [37]. Successful tests of RuNet independence were reported in 2019 and 2021 [38].

Subsequent to the Russian invasion of Ukraine in February 2022, many foreign service providers withdrew from the Russian market [39]. Others were banned by Russia. The Ukrainian representative at ICANN requested that top-level Russian domains and certificates be revoked by ICANN [40]. ICANN refused this unprecedented request. In early March 2022, the rumour that Russia would disconnect itself entirely from the global Internet on March 11, apparently based on a Kremlin document on preparing for separation, was widespread [41]. Some commentators suggested that this would presage an all-out cyberattack on the U.S., or the cutting of transatlantic cables by Russia. Calls for Russia to be disconnected from the Internet, and rumours that it will disconnect itself, are still circulating at time of writing in April 2022.

III. SECURITY ISSUES FOR A GLOBAL INTERNET

Metcalfe's law states that the value of a communications network is proportional to the square of the number of connected users of the system [50]. However, it has been shown that an increase in the number of connected users also increases risk, which in turn diminishes value [38]. In this paper we argue that the uncertainty and fragility caused by widespread insecure software is likely to add further pressure to a global Internet infrastructure that is already fragmenting. We base this argument on the fact that insecure software facilitates crime, espionage and sabotage across borders. In this section we discuss the top crimes and threats from the Threat Landscape report issued by The European Union Agency for Cybersecurity (ENISA) [42], which covers the year prior to July 2021. Published in October 2021, the report lists the main threats encountered and defines four categories of threat actor. Like the ENISA report, we do not consider localised issues such as those related to intimate partner abuse and cyberbullying, because those very real risks are not primarily international. Having discussed threats defined by ENISA, we add cyber patriotism and cyberwarfare.

A. Threat actors in the ENISA report

The ENISA report defines four different threat actors.

1) State-Sponsored Actors: The report (see Table I) describes a rise in cyberespionage related to Covid-19, with state actors observed searching for information on national Covid-19 responses and treatment. Healthcare and medical research sources were targeted. Supply-chain compromises were significant, in particular the highly sophisticated SolarWinds SunBurst breach [43]. State actors were observed engaging in money-making activities such as cryptojacking, perhaps partially to disguise breaches as cybercrime.

Both defenders and state actors raised their game in the reporting period, with numerous joint declarations and legal stratagems. State actors showed increasing levels of sophistication. 'False flags' were sometimes used to muddy attribution, and hack-and-leak campaigns were used for strategic gains.

2) Cybercriminals: Covid-19 was used by cybercriminals in multiple phishing campaigns preying on concern about the virus. The report notes increased collaboration and professionalism, a move to the cloud and an increasing tendency to attack critical infrastructure. The report mentions the 'Cybercrime-as-a-Service' trend, wherein services for cybercrime are commoditised and broken down; it is possible to purchase access to victim servers from one dark web supplier and run ransomware on them which has been purchased from a different supplier. Many other services are offered in this ecosystem. Since it is global, hackers in one country can sell their services to cybercriminals in another.

3) Hacker-For-Hire Actors: The ENISA report described the Access-as-a-Service (AaaS) market. Commonly known as spyware, AaaS allows the user to access the contents of a victim's phone, potentially including the microphone and camera. The report predicted that this sector will be subject to increasing regulation on human rights as well as national security grounds. This prediction has been borne out by events. In November 2021, the U.S. blacklisted well-known AaaS firm NSO group [44], and Israel drastically reduced the number

of countries to which cyber-weapons could be exported [45]. The technology continued to cause controversy in 2022, with a stream of revelations including the discovery in February that Israel had used NSO spyware against some of its own public figures [46]. In April, use of NSO spyware for surveillance of Jordanian human rights defenders was revealed [47].

4) Hacktivists: Early hacktivism was generally associated with idealistic left-wing anti-corporate ideology. Hacktivists use cyberspace for activities related to political activism in the real world, aiming to increase awareness or to cause reputational damage to organisations. Hacktivism is typically not done for financial or material gain [48]. The ENISA report finds low current levels of hacktivism, but anticipates a possible rise in the future as environmental issues come to the fore. It notes that hacktivism can be faked by nation state actors to confuse attribution for subversive activities.

In October 2021, protests in Belarus over the disputed re-election of Alexander Lukashenko were accompanied by hacktivist activity, including the theft and release of information revealing the identities of Belarussian security agents [49].

B. Cybercrime threats in 2020-2021

1) Ransomware: Ransomware is the practice of encrypting the files on an organisation's devices and demanding a ransom for the decryption key. Since CryptoLocker first appeared in September 2013 [50], ransomware has become increasingly sophisticated. Recent escalation tactics include using Distributed Denial of Service (DDoS) [51], and threatening to expose sensitive data, including embarrassing data from the devices of organisational decision-makers [52]. Some hackers search networks for details of cybersecurity insurance coverage amounts, tailoring their ransom requests accordingly [50]. With an estimated $590 million of ransomware payments made in the first six months of 2021 [53], by November an insurance backlash had begun, with rises in premiums of up to 300% and steep falls in amounts covered [54].

Ransomware crews make their expertise available to franchisees in what is known as a Ransomware-as-a-Service model [55]. They take precautions to ensure that franchisees do not launch attacks in their home countries, often automating a check of the installed language on a system before file encryption [56]. Security journalist Brian Krebs suggested that installing certain Eastern European languages on a computer could provide protection against some ransomware strains [57].

The ENISA report describes how zero-day vulnerabilities, generally bought by nation-state actors, were in 2021 often used in sophisticated attacks on small numbers of very high-value ransomware targets, a practice known as big game hunting.

In June 2021, the U.S. government raised the priority of ransomware to the same level as terrorism [58]. Subsequent initiatives such as the international 'Counter Ransomware Initiative' [59] sought to improve international ransomware prevention and response. Priorities were increasing resilience, disrupting illicit finance and jurisdictional arbitrage, and improving international cooperation and diplomacy to encourage states to address ransomware operations within their own territories [60]. A series of arrests and forum shutdowns by Russian authorities in January and February 2022 was considered a change in Russian policy towards ransomware and other cybercrime [61]. Whether a conciliatory gesture towards the U.S. [62], or an attempt to keep China, also experiencing severe ransomware incursions [63], onside, enforcement diminished after Russia invaded Ukraine.

Industry commentators in early 2022 observed increased use of ransomware by nation state actors, such as a January fake ransomware attack on Ukrainian government sites, concluding that the ransomware cover provides deniability to an attacking state [64].

2) Cryptojacking: Often seen as a relatively victimless crime, cryptojacking is the practice of surreptitiously mining cryptocurrency on a user's device. When done at scale it can be lucrative [65]. ENISA reports that cryptojacking incidence was at its highest ever in the first quarter of 2021. It suggests that the rapid increase of cryptojacking and ransomware is facilitated by the ease with which they translate to financial gain, facilitated by the use of cryptocurrencies.

3) Other Cybercrime: While cryptojacking and ransomware require large-scale networks to function, there is also plenty of traditional crime on the Internet. In an analysis of the 'Digital Goods' or 'Services' dark web sales categories, Meland et al. [55] report that credit card fraud ('carding') is the most popular crime. Carding involves the bulk selling of credit card data, sometimes with card holders' personal details [66]. Stealing and selling credit card information at scale is easier online.

4) Cyberespionage: Cyberespionage is now an accepted part of geopolitics. Between December 2020 and February 2021, national infrastructure cyber-intrusions were reported by Finland (parliamentary email) [67], Japan (military contractor) [68], Malaysia (Armed Forces website) [69] and Ukraine (government document sharing) [70], to take just a few examples. In early 2021, intrusions on U.S. and other government networks via Sunburst (SolarWinds) and other supply chain attacks caused concern about the risk of cyberespionage. However, experts in the field expressed the view that this was merely traditional international jostling [43].

5) Cyber Patriotism: Not mentioned in the ENISA report, which concluded observations in mid-2021, there has been an outbreak of activity from what Recorded Future's Allan Liska calls 'cyber patriots' as a result of Russia's invasion of Ukraine. We distinguish cyber patriotism from hacktivism on the basis of its nationalist origins. Sharp divisions have occurred within cybercriminal groups that contained members from both Russia and Ukraine [71]. Many hacker groups have taken sides, vowing to leverage their skills to further their country's cause [72]. Others have also acted. After the notorious Conti ransomware group announced its support for Russia, a Ukrainian researcher, who had lurked on Conti servers for years, leaked thousands of documents containing their internal communications [73].

Cyber patriotism has had a direct impact on software security. Some software component projects on GitHub have been modified to become 'protestware,' displaying banners like 'Stand with Ukraine,' or facts about the invasion. In one case,

the popular 'vue-cli' framework had a component added that deleted all files on its host computer if it detected that it was running in Russia or Belarus. Brian Krebs reports concerns that such activities would '*erode public trust in open-source software*' [74]. Since blind trust in open source components is not conducive to software security, we argue that this might be a good thing.

6) Cyberwarfare: As it moves online, infrastructure is increasingly vulnerable to cyber outages. These can be caused by natural phenomena such as hurricanes. They can be collateral damage from criminal cyber activity, as the Colonial pipeline outage in the U.S. in May 2021 was. They can also be the result of actions by a hostile state. Cybersecurity organisation Recorded Future documented a large increase in suspected intrusion activity in India by Chinese state-sponsored groups during border tensions in 2020. Recorded Future stated that India's power sector and two seaports were targeted in a *'concerted campaign against India's critical infrastructure.'* Severe power outages in Mumbai on October 12 2020 were attributed to Chinese sabotage by Anil Deshmukh, a minister for Maharashtra state. China disputes the claim, but the fact that it was made at all reflects the uncertainty engendered by the mere possibility of cyberattack.

In 2010 the Stuxnet worm, widely attributed to Israel and the U.S., attacked industrial control systems in Iran. The Natanz uranium enrichment site was badly damaged, even though the Natanz network was supposedly air-gapped from the Internet. Stuxnet is considered to be the world's first cyber-weapon [75].

Prior to the February 2022 invasion of Ukraine, Russia-Ukraine history showed a gradual escalation from cyberwarfare to kinetic warfare. The electric grid in Ukraine was attacked on December 23rd 2015. In an incursion attributed by the DoJ to Russia's GRU [76], 30 substations were taken offline and power to 230,000 people in freezing temperatures was lost for up to 6 hours. There were related outages the following year. In 2017, an accounting tool used by approximately half of the businesses in Ukraine was infiltrated with fake ransomware in what became known as the NotPetya attack. There were huge financial costs to business. The Merck pharmaceutical company lost $1.4bn [77]. This incident was attributed to the Russian state by the UK government [78], but the apparently criminal method of attack allowed for plausible deniability. It was hugely destabilising in Ukraine, and served as a warning to international organisations considering doing business there, signalling that perhaps it would not be worth the trouble [79]. In 2020, the DoJ indicted six Russian nationals for the Ukrainian power cuts and the NotPetya attack, among other alleged crimes [76]. An unintended victim was the insurance industry, forced to contend with geopolitical questions around attribution and 'act of war' definitions in its attempts to avoid payouts [77].

In the build-up to the Russian invasion, cyberattacks on Ukraine increased, with data wipers disguised as ransomware [80], DDoS, bot farms spreading misinformation [81] and widespread infrastructure attacks [82]. A cyberattack on the day of the invasion on Viasat KA-SAT routers used in Ukrainian military communications had an impact on other European countries, with monitoring and control of wind turbines in Germany rendered unavailable [83]. Cyberattacks continued after the invasion [82]. Meanwhile, western officials warned amateur hackers against joining the voluntary 'IT Army of Ukraine,' organising on Telegram.

In a discussion on cybersecurity threat escalation on the website of the Arms Control Association (ACA), Michael T. Klare describes the inherent danger that a cyberattack on Nuclear Command, Control, and Communications (NC3) facilities would justify a nuclear response. The ACA views this as an unacceptable risk, suggesting that even the fear that NC3 facilities were under attack could trigger an escalation to the use of nuclear weapons. If tensions were high enough, even a simple power outage could cause a national leader to feel that their nuclear capability was in imminent danger. This could propel them into striking first [84]. The advent of cyber patriot vigilantes, some of them expert hackers, increases the risk of such an unanticipated outcome.

The danger that cyber incidents could cause escalation to kinetic warfare was raised by U.S. President Biden in 2021 [85]. It is likely to be considered by every nation when assessing the pros and cons of unfettered access to a global Internet.

IV. The Role of Software Vulnerabilities

Thus far, we have discussed some of the forces pushing nations to separate from the global Internet, and outlined some of the threats that are likely to accelerate that process. We now turn to the role of software vulnerabilities in exacerbating those threats. A perusal of material from hacker training guides such as the Web Application Hacker's Handbook [86] indicates that discovery and use of software vulnerabilities is at the core of hacking techniques.

Vulnerabilities are mitigated by software patches. In a survey by *BAE Systems Applied Intelligence*, reported in May 2021 [87], 52% of recent security incidents were caused by missing patches. The mean time to patch was 205 days [88]. Patching strategies are complicated by the composition of modern software, which normally contains multiple open source software (OSS) components which may themselves include other libraries. One third of studied vulnerabilities in OSS were present for over three years before remediation [89]. December 2021 brought this issue to the fore with the publicising of the Log4j bug, in which a little-known feature of a ubiquitous Java logging component was discovered to be vulnerable to remote code execution [90].

Unfortunately, vulnerable systems are easily discoverable online. Actors wishing to exploit the latest defects can run tailored searches via sites such as Shodan [91], which will find and list Internet-facing systems with specified characteristics. Failure to patch is discoverable.

Not all software vulnerabilities are equal. In the U.S., NIST maintains the National Vulnerability Database, which collates reported software vulnerabilities and assigns a Common Vulnerability Scoring System (CVSS) score to them, with a 'Critical' 10.0 being the highest score available. High CVSS scores indicate that a vulnerability is simple to exploit, remotely

available, and likely to result in a severe impact on the vulnerable system. The term 'zero-day' is used to describe a critical software vulnerability that has not yet been patched, may not be generally known about, and possibly has not even been reported to the software manufacturer. Zero-days for popular software are much in demand and can be bought on the dark web [92]. National security agencies are known to stockpile zero-days for use in cyberespionage [93].

In April 2017, the 'Shadow Brokers' published a number of hacking tools widely reputed to originate with the U.S. National Security Agency (NSA) [94]. These leveraged serious bugs such as the Eternal Blue exploit (CVE-2017-0144), which the NSA had reputedly used for several years [93]. Microsoft had been notified of the theft of the exploit, and released a patch for Eternal Blue a month before it was published [95]. Nevertheless, sufficient machines remained unpatched for the WannaCry ransomware cryptoworm attack of May 12 2017 and the NotPetya attack of June 27 2017 to cause worldwide havoc. The Eternal Blue SMB exploit allowed WannaCry and NotPetya to spread and self-propagate without any user intervention [96], [97]. Until it was patched, Eternal Blue was present on all versions of Windows from at least Windows 2000.

Another long-lived critical Microsoft defect was the 'Zerologon' elevation-of-privilege bug. Quietly patched in August 2020 and made public the following month, it infected Windows Server 2008 and all newer versions of Windows Server up to 2019 [98]. The persistence of Eternal Blue and Zerologon for over a decade after Microsoft mandated internal use of Microsoft Security Development Lifecycle (MS-SDL) is a reminder that there are no silver bullets when it comes to software security.

The relatively collegiate international atmosphere that had surrounded defect discovery and notification began to change in 2018, when the Chinese government banned Chinese security researchers from participating in vulnerability discovery competitions such as CanSecWest's Pwn2Own [99], in which they had previously been highly successful. In 2021, the Cyberspace Administration of China introduced rules forbidding the sale of vulnerabilities or the notification of vulnerabilities to overseas entities other than the manufacturers. Organisations discovering vulnerabilities in their own code must notify them to the Chinese government within two days [100]. For entities trading within China, this could put them in a position of having to notify the Chinese government about vulnerabilities before a patch is in place. Organisations are 'encouraged,' though not obliged, to notify the government first about vulnerabilities discovered in other organisations' code. In December 2021, AliBaba Cloud was suspended from an information-sharing partnership with China's Ministry of Industry and Information Technology (MIIT) for failure to notify it about the Log4j vulnerability. AliBaba staff notified Apache on November 24, while the MIIT was not notified until December 9 [101].

Considered in light of the move by some countries to use only homegrown software internally, these developments could presage a time when foreign adversaries are familiar with the software used by the U.S. and the EU, and its vulnerabilities,, while the reverse is no longer true.

V. Attempts to Remediate

Having seen the impact of software vulnerabilities on software quality, we now look at approaches in industry and academia to reducing software vulnerabilities. We consider some of the shortcomings of existing approaches and suggest some reasons why they are not effective. We also discuss recent legislation relating to software security in Europe and the U.S.

A. Industry

Focus on software security in industry varies depending on the industry involved. In the U.S., NIST publishes comprehensive cybersecurity guidelines. Revision 5 of NIST Special Publication 800-53, 'Security and Privacy Controls for Information Systems and Organizations' was published in September 2020 [102]. The guide is used by safety-critical industries; for example, U.S. Nuclear Regulatory Commission Regulatory Guide 5.71, on cybersecurity programmes for nuclear facilities, used NIST 800-53 version 3 to provide a comprehensive cybersecurity approach [103]. Weighing in at a hefty 465 pages, version 5 provides descriptions of numerous cybersecurity controls but includes a mere four pages on 'Developer Testing and Evaluation.' This software development section outlines nine activities that would be familiar to most software security advocates.

Software security methodologies in general use in industry include MS-SDL, Software Assurance Maturity Model, Building Security In Maturity Model, Common Criteria and various ISO standards. All coalesce around a number of activities which are regularly synthesised in academic papers [104] [105], such as threat modelling, use of analysis tools and penetration testing. However, the software development security industry is currently convulsed by software developers' move away from regular, relatively infrequent releases, to which blocking 'security gates' can be applied, to automated continuous releases. This move is facilitated by the DevOps emphasis on comprehensive automated tests. DevSecOps attempts to bring security into the DevOps approach, adding security tests to the automated test suite and automating security gates. Advocates of DevSecOps suggest that it 'shifts security to the left,' making it an issue that architects and developers must consider instead of something that is assessed just before release. Done well, DevSecOps can add predictability and credibility to a development team's security stance. However, practitioners express concerns about whether comprehensive security checks can ever be fully automated [106]. Done badly, DevSecOps adds little value and can even have a negative impact on a development process [107].

B. Academia

Much work has been done in academia on how software can be made more secure. Wurster and von Oorschot [108] argued that software developers, though seen as warriors in the forefront of the battle for secure software, are in fact part of the problem since they have multiple, often conflicting,

priorities and are rarely security experts. They suggested that developer tools should be created with usability in mind, and should make it difficult for developers to code insecurely. They pointed out that developer security training was often still advocated as the solution for developers. They identified an issue with developers who are either unaware of, or who ignore, new security technologies, and noted that security technologies which must be independently run by developers (i.e. they are not embedded in standard tools) will not be run by all developers. They advocated 'security mechanisms which are invisible to the application developer.' This theme was developed by Xie et al. [109], who looked at why programmers make security errors, concluding that developers often feel that someone else is responsible for security and it is not their concern. Acar et al. [110] discussed how 20 years of lessons learned from usable security work can be applied in security research with software developers, and derived a research agenda on these lines. Green and Smith [111] discussed simplifying security APIs to make them less impenetrable to programmers, proposing ten principles for creating secure and usable crypto APIs.

A recurring research theme is that developers, who have other priorities, lack the training and expertise necessary for security proficiency. Weir et al. [112], the Motivating Jenny team (https://motivatingjenny.org/) and others have looked at interventions and tools to help teams to code securely. However, in a 2020 review of top U.S. Computer Science (CS) undergraduate courses, Almansoori et al. [113] found that security-unaware use of insecure C++ functions was passed from teachers to students. Moreover, in all cases there was no mandatory formal secure coding component to the CS course. We argue that while *ad hoc* on-the-job training efforts have value, software security is so critical that it should be automatically embedded in all software development training.

The academic record includes valuable accounts of actual industry practice. Sadowski et al. [114] described the development of a static analysis tool at Google, which is a model for what can be done in a cohesive environment, even a very large one. By contrast, Morales et al.'s [107] account of a dysfunctional multi-year development by a main contractor using multiple subcontractors with a DevSecOps pipeline gives excellent insight into the ease with which organisational dynamics can damage security outcomes. A review of the literature suggests that swift software security progress is likely to be tied to upper management concern, which may be enhanced by increased regulatory and legal incentives.

C. Legislation

Both Europe and the U.S. appear to be moving towards regulations for secure software, a welcome recognition of the increasing importance of the field. Here, we give a brief overview of relevant developments.

GDPR: The EU's General Data Protection Regulation (GDPR), which governs the protection of personal data in the European Economic Area, came into force in 2018. It mandates obtaining subjects' permission for data storage, sets time limits on data retention, and allows for penalties where data is inappropriately shared. Although the GDPR was not created with the primary aim of enhancing software security it has had this effect, since a data breach could expose the organisation to financial penalties. It does not list explicit cybersecurity requirements, instead using broad phrases such as 'state of the art.' This deters a 'checkbox' security mentality, encouraging awareness of ongoing software security developments [115].

The NIS Directive and ENISA: The EU's 2016 NIS Directive dealt with cybersecurity but did not discuss secure software development [116], though this should change with NIS2. A thoughtful and well-researched preparatory paper from ENISA outlines the current EU work on introducing security certification for software, with consideration of existing standards and certifications, and likely pitfalls [117]. This welcome move towards EU-level certification should be expedited.

The UK: The UK's 'Government Cyber Security Strategy 2022-2030,' released in February 2022, lists aspirations around a 'secure by design' framework to be adopted by the UK [118]. No specific advice for software development is available yet.

The U.S.: In a week in May 2021 in which the ransomware shutdown of an essential U.S. oil pipeline dominated the news, the U.S. President released an 'Executive order on Improving the Nation's Cybersecurity,' which provides specific software-related measures. Part a) asserts that *'the Federal Government must take action to rapidly improve the security and integrity of the software supply chain, with a priority on addressing critical software.'* NIST is required to identify guidelines to evaluate software security and the security practices of developers and suppliers, and to identify *'tools or methods to demonstrate conformance with secure practices.'* Enhancing supply chain security, securing build environments, automating supply chain assurance and providing evidence for these activities are discussed. Identifying 'critical' software is also addressed, as is Internet of Things (IoT) security and a suggested security labelling system for software and IoT devices. U.S. government agencies will be obliged to consider software security when engaging in or renewing critical software contracts. Legacy code that cannot comply with the new requirements will have to be replaced. Steps to secure the software supply chain will be kept under review, with a progress report required within a year of the signing of the order.

D. Exacerbating Factors

There is an essential imbalance in the software security world. Most software developers prioritise functionality [119], followed by efficiency [106], elegant design, or maintainability. Unless they are working for an organisation that emphasises security, they will probably not put security first. In fact, even if they work for a security organisation, their security practice could be suspect [120].

While software developers struggle with time-to-market and tight deadlines, hackers, security researchers and red-teamers can focus solely on finding the security defects inadvertently left by developers, and exploiting them. When it comes to training, they have multiple resources at their disposal such as the free Bugcrowd University

(https://www.bugcrowd.com/hackers/bugcrowd-university) and the many dozens of courses annually at Black Hat and elsewhere geared towards 'penetration testers'.

This imbalance of time and resources is difficult to tackle. Many software developers have not received training in secure coding. Organisations often have relatively few, if any, software security staff; a single software security expert supporting one or two hundred developers is not uncommon [121]. Outside of regulated industries, there is currently little incentive for organisations to prioritise security over time-to-market, and some organisations have no security process at all [104].

E. A 'Zero Tolerance' Approach to Software Security

We argue that the software industry now needs to step up and adopt a 'zero tolerance' approach to software security issues. Haney et al. [122] described how organisations that successfully deliver secure software have a 'security culture.' The entire software industry needs to develop a 'security culture,' with a comprehensive upgrade of education, tools, and documentation.

The building blocks to achieving this are not novel. They involve steps that are both widely acknowledged as necessary, and widely ignored. Cultural change is needed in education, from universities to boot camps. It should be unacceptable to teach computer skills without including mandatory security awareness and associated training.

At the corporate level, too often security risk assessments end with a decision to increase insurance provision against cyberattack. The balance of risk must be changed. This can be done by making organisations liable for costs incurred due to secure coding negligence on their part. Some experts argue that mandating secure coding would impose the type of procedural rigidity that leads to an obsession with passing tests, as can happen in the payments industry [123]. However, the flexible wording of the GDPR gives an insight into how secure coding can be mandated without leading to a checkbox mentality. In any case, even a checkbox mentality would be a vast improvement on the current security posture of many organisations [124].

VI. Conclusion

As we have seen, crime and other aggressive behaviour on the Internet thrives on the existence of software vulnerabilities. A steady supply of zero-days, combined with delays in patching known vulnerabilities, ensures that bad actors can continue to exploit weaknesses for financial or other gain. This state of affairs causes an unsustainable level of uncertainty and risk. The threat of industrial sabotage or breach of military command and control structures from foreign actors adds to the mounting pressures on the global Internet, pushing it towards further fragmentation.

We have also seen how the opportunities for incursion provided by a global Internet can have a destabilising effect on existing power balances. If software security is not taken sufficiently seriously, there is a danger that national administrations will increasingly judge that the price of participating in a global network is too high. National executives may even decide that retreating to a national or regional Intranet would enhance their national security and reduce the danger of cyberattack on NC3 facilities and other critical infrastructure. As incidents of damage from cyber activities increase globally, assessments of this type may not be confined to authoritarian regimes. Though some states might welcome a fragmentation of the Internet, it seems like a failure of human imagination and potential.

The complete elimination of software vulnerabilities may be impossible, but a drastic reduction is not. It is time for a less accommodating approach. A 'zero tolerance' attitude to software security issues should be adopted, and it should include cultural and legislative change. This would reduce the perceived vulnerability of vital systems and help to maintain confidence in a networked world.

Acknowledgment

SFI grants 18/CRT/6222, 13/RC/2077_P2, 13/RC/2094_P2.

References

[1] S. Hoffmann, D. Lazanski, and E. Taylor, "Standardising the Splinternet: how China's technical standards could fragment the Internet," *J. Cyber Policy*, vol. 5, 2020.
[2] S. Harrison, "Cyber attack: When will the Irish health service get a resolution?" https://www.bbc.com/news/world-europe-57193160, 2020.
[3] R. Lakshmanan, "Exclusive: SonicWall hacked using 0-day bugs in its own VPN product," https://thehackernews.com/2021/01/exclusive-sonicwall-hacked-using-0-day.html, 2021.
[4] B. Dickson, "AV Oracle: New hacking technique leverages antivirus to steal secrets," https://portswigger.net/daily-swig/av-oracle-new-hacking-technique-leverages-antivirus-to-steal-secrets, 2019.
[5] M. Korolov, "Cisco router vulnerability puts network segmentation at risk," https://www.datacenterknowledge.com/security/cisco-router-vulnerability-puts-network-segmentation-risk, 2020.
[6] A. O'Driscoll, "25+ cyber security vulnerability statistics and facts of 2021," https://www.comparitech.com/blog/information-security/cybersecurity-vulnerability-statistics/, 2021.
[7] E. Claessen, "Reshaping the internet – the impact of the securitisation of internet infrastructure on approaches to internet governance: the case of Russia and the EU," *J. Cyber Policy*, vol. 5, no. 1, 2020.
[8] D. Dominey-Howes, "The Tonga volcanic eruption reveals the vulnerabilities in our global telecommunication system," https://techxplore.com/news/2022-01-tonga-volcanic-eruption-reveals-vulnerabilities.html.
[9] E. Buchanan, "Subsea cables in a thawing Arctic," https://www.maritime-executive.com/editorials/subsea-cables-in-a-thawing-arctic, 2018.
[10] C. Gallagher, "Russian military drills pose strategic and environmental risks to Ireland," https://www.irishtimes.com/news/ireland/irish-news/russian-military-drills-pose-strategic-and-environmental-risks-to-ireland-1.4784787.
[11] K. Ermoshina, B. Loveluck, and F. Musiani, "A market of black boxes: The political economy of Internet surveillance and censorship in Russia," *Journal of Information Technology & Politics*, 2021.
[12] FCC, "Circuit capacity data for U.S.-international submarine cables," https://www.fcc.gov/international/circuit-capacity-data-us-international-submarine-cables, 2022.
[13] J. Smith, "Internet Atlas maps the physical internet to enhance security," https://news.wisc.edu/internet-atlas-maps-the-physical-internet-to-enhance-security/, 2021.
[14] D. Temple-Raston, "Report: Beijing, Moscow step up efforts to control the Internet's backbone," https://therecord.media/report-beijing-moscow-step-up-efforts-to-control-the-internets-backbone/, 2021.
[15] H. Fouquet, "China's 7,500-mile undersea cable to Europe fuels Internet feud," https://www.msn.com/en-us/money/other/china-e2-80-99s-7500-mile-undersea-cable-to-europe-fuels-internet-feud/ar-BB1egCN9.
[16] J. Sherman, "The US-China battle over the Internet goes under the sea," https://www.wired.com/story/opinion-the-us-china-battle-over-the-internet-goes-under-the-sea/, 2020.
[17] M. Cartwright, "Internationalising state power through the Internet: Google, Huawei and geopolitical struggle," *Internet Policy Review*, vol. 9, no. 3, 2020.

[18] J. Dunleavy, "FCC designates Huawei and four other Chinese tech companies as national security threats," https://www.washingtonexaminer.com/news/fcc-huawei-four-other-chinese-tech-companies-national-security-threats, 2021.

[19] Cloudflare, "What is an Internet exchange point? — How do IXPs work?" https://www.cloudflare.com/learning/cdn/glossary/internet-exchange-point-ixp/, 2019.

[20] D. McCullagh, "How Pakistan knocked YouTube offline (and how to make sure it never happens again)," https://www.cnet.com/news/how-pakistan-knocked-youtube-offline-and-how-to-make-sure-it-never-happens-again/, 2008.

[21] V. Ververis, S. Marguel, and B. Fabian, "Cross-country comparison of Internet censorship: A literature review," *Policy Internet*, vol. 12, no. 4.

[22] K. Limonier, F. Douzet, L. Pétiniaud, L. Salamatian, and K. Salamatian, "Mapping the routes of the Internet for geopolitics: The case of Eastern Ukraine," *First Monday*, 2021.

[23] W. J. Drake, V. G. Cerf, and W. Kleinwächter, "Internet fragmentation: An overview," p. 80, 2016.

[24] Z. Chen, C. Wang, G. Li, Z. Lou, S. Jiang, and A. Galis, "NEW IP framework and protocol for future applications," in *NOMS 2020*.

[25] R. D. Taylor, ""Data localization": The Internet in the balance," *Telecommunications Policy*, vol. 44, no. 8, 2020.

[26] J. Silva, "LinkedIn is shutting down in China, will be replaced by a new app called InJobs," https://www.techspot.com/news/91754-linkedin-shutting-down-china-replaced-new-app-called.html, 2021.

[27] R. Harb, "India bans a further 118 Chinese apps as physical and online tensions escalate," https://www.theregister.com/2020/09/03/india_bans_chinese_apps/.

[28] D. Goodin, "FCC puts Kaspersky on security threat list, says it poses "unacceptable risk"," https://arstechnica.com/information-technology/2022/03/fcc-puts-kaspersky-on-security-threat-list-says-it-poses-unacceptable-risk/.

[29] F. House, "Countries - Internet freedom scores," https://freedomhouse.org/countries/freedom-net/scores, 2021.

[30] A. A. Niaki, S. Cho, Z. Weinberg, N. P. Hoang, A. Razaghpanah, N. Christin, and P. Gill, "ICLab: A global, longitudinal Internet censorship measurement platform," in *IEEE SP 2020*.

[31] B. Duggan, "Uganda shuts down social media; candidates arrested on election day," https://edition.cnn.com/2016/02/18/world/uganda-election-social-media-shutdown/index.html.

[32] HRW, "Belarus: Internet disruptions, online censorship," https://www.hrw.org/news/2020/08/28/belarus-internet-disruptions-online-censorship, 2020.

[33] BBC, "Kazakhstan unrest: Internet cut amid fuel protests," https://www.bbc.com/news/world-asia-59876093, 2022.

[34] C. Cimpanu, "Oracle: China's internet is designed more like an intranet," https://www.zdnet.com/article/oracle-chinas-internet-is-designed-more-like-an-intranet/, 2019.

[35] ——, "Two of China's largest tech firms are uniting to create a new 'domestic OS'," https://www.zdnet.com/article/two-of-chinas-largest-tech-firms-are-uniting-to-create-a-new-domestic-os/, 2019.

[36] "Russia to launch 'independent internet' for BRICS nations - report," https://www.rt.com/russia/411156-russia-to-launch-independent-internet/.

[37] BBC, "Russia bans sale of gadgets without Russian-made software," https://www.bbc.com/news/world-europe-50507849, 2019.

[38] S. Jarman, "How pulling out of Russia's internet could further isolate its citizens," https://bigthink.com/the-present/russian-internet-runet/?utm_medium=Social&utm_source=Twitter#Echobox=1648510641-2.

[39] H. F. Ukraine, "Companies suspending operations in Russia and Belarus," https://hrforukraine.notion.site/Companies-Suspending-Operations-in-Russia-and-Belarus-93d42cbb7a234438b663bded9f1f7f4c.

[40] S. Fadilpašić, "ICANN rejects call to remove Russian domains from the Internet," https://www.msn.com/en-us/news/technology/icann-rejects-call-to-remove-russian-domains-from-the-internet/ar-AAUBABO.

[41] J. Parsons, "Russians 'to be disconnected from global internet from Friday'," https://metro.co.uk/2022/03/07/russia-preparing-to-disconnect-from-global-internet-on-march-11-16230918/.

[42] EU, *ENISA threat landscape 2021: April 2020 to mid July 2021*. Publications Office.

[43] T. Wheeler, "The danger in calling the SolarWinds breach an 'act of war'," https://www.brookings.edu/techstream/the-danger-in-calling-the-solarwinds-breach-an-act-of-war/, 2021.

[44] BBC, "NSO Group: Israeli spyware company added to US trade blacklist," https://www.bbc.com/news/technology-59149651, 2021.

[45] C. Cimpanu, "Israel restricts cyberweapons export list by two-thirds, from 102 to 37 countries," https://therecord.media/israel-restricts-cyberweapons-export-list-by-two-thirds-from-102-to-37-countries/.

[46] Reuters, "Israel ramps up scrutiny of police as NSO scandal spreads," https://www.euronews.com/2022/02/07/us-israel-nso.

[47] FLD, "Report: Jordanian human rights defenders and journalists hacked with Pegasus spyware," https://www.frontlinedefenders.org/en/statement-report/report-jordanian-human-rights-defenders-and-journalists-hacked-pegasus-spyware.

[48] A. Pawlicka, M. Choraś, and M. Pawlicki, "Cyberspace threats," in *ARES '20*.

[49] O. Carroll, "Hacktivists vs The Dictator: How Belarus cyber army is taking on Alexander Lukashenko and his goons," https://www.independent.co.uk/news/world/europe/belarus-lukashenko-protests-cyber-attacks-minsk-b807184.html, 2021.

[50] S. Greengard, "The worsening state of ransomware," *Commun. ACM*, vol. 64, no. 4, Apr 2021.

[51] D. S. Amanda Tanner, Alex Hinchliffe, "Threat assessment: Blackcat ransomware," https://unit42.paloaltonetworks.com/blackcat-ransomware/.

[52] C. Cimpanu, "Some ransomware gangs are going after top execs to pressure companies into paying," https://www.zdnet.com/article/some-ransomware-gangs-are-going-after-top-execs-to-pressure-companies-into-paying/, 2021.

[53] "Treasury continues to counter ransomware as part of whole-of-government effort; sanctions ransomware operators and virtual currency exchange," https://home.treasury.gov/news/press-releases/jy0471.

[54] C. Cohn, "Insurers run from ransomware cover as losses mount," https://www.reuters.com/markets/europe/insurers-run-ransomware-cover-losses-mount-2021-11-19/, 2021.

[55] P. H. Meland, Y. F. F. Bayoumy, and G. Sindre, "The Ransomware-as-a-Service economy within the darknet," *Computers & Security*, vol. 92.

[56] C. Nocturnus, "Cybereason vs. DarkSide Ransomware," https://www.cybereason.com/blog/cybereason-vs-darkside-ransomware.

[57] B. Krebs, "Try this one weird trick Russian hackers hate," https://krebsonsecurity.com/2021/05/try-this-one-weird-trick-russian-hackers-hate, 2021.

[58] M. Sharma, "US will give ransomware hacks similar priority to terrorist attacks," https://www.techradar.com/uk/news/us-will-give-ransomware-hacks-similar-priority-to-terrorist-attacks.

[59] A. Neuberger, "Update on the International Counter-Ransomware Initiative," https://www.state.gov/briefings-foreign-press-centers/update-on-the-international-counter-ransomware-initiative.

[60] T. SpiderLabs, "Law enforcement collaboration has Eastern-European cybercriminals questioning whether there is a safe haven anymore," https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/law-enforcement-collaboration-has-eastern-european-cybercriminals-questioning-whether-there-is-a-safe-haven-anymore/, 2021.

[61] A. Vicens, "Russian government continues crackdown on cybercriminals," https://www.cyberscoop.com/sky-fraud-takedown-russia-cybercrime/.

[62] M. I. Nicole Sganga, "Russia arrests 14 alleged members of REvil ransomware gang," https://www.cbsnews.com/news/ransomware-russia-arrests-revil/, 2022.

[63] T. Uren, "Srsly risky biz: Thursday January 13," https://srslyriskybiz.substack.com/p/srsly-risky-biz-thursday-january-39c.

[64] C. Team, "As ransomware payments continue to grow, so too does ransomware's role in geopolitical conflict," https://blog.chainalysis.com/reports/2022-crypto-crime-report-preview-ransomware/.

[65] S. Pastrana and G. Suarez-Tangil, "A first look at the crypto-mining malware ecosystem: A decade of unrestricted wealth," in *IMC '20*.

[66] P.-Y. Du, N. Zhang, M. Ebrahimi, S. Samtani, B. Lazarine, N. Arnold, R. Dunn, S. Suntwal, G. Angeles, R. Schweitzer, and et al., "Identifying, collecting, and presenting hacker community data: Forums, IRC, carding shops, and DNMs," in *ISI 2018*.

[67] C. Cimpanu, "Finland says hackers accessed MPs' emails accounts," https://www.zdnet.com/article/finland-says-hackers-accessed-mps-emails-accounts/, 2020.

[68] "Kawasaki Heavy hack may have targeted defense-linked information," https://proteuscyber.com/privacy-database/news/3014-kawasaki-heavy-hack-may-have-targeted-defense-linked-information-the-japan-times.

[69] "Malaysia's armed forces confirms cyber-attack on network," https://www.straitstimes.com/asia/se-asia/malaysias-armed-forces-confirms-cyber-attack-on-network, 2020.

[70] "The NCCC at the NSDC of Ukraine warns of a cyberattack on the document management system of state bodies," https://www.rnbo.gov.ua/en/Diialnist/4823.html, 2021.

[71] J. Uchill, "After Conti backs war, ransomware gangs realize peril of patriotism amid infighting," https://www.scmagazine.com/analysis/ransomware/after-conti-backs-war-ransomware-gangs-realize-peril-of-patriotism-amid-infighting.

[72] E. Vail, "Russia or Ukraine: Hacking groups take sides," https://therecord.media/russia-or-ukraine-hacking-groups-take-sides/.

[73] S. Lyngaas, "'I can fight with a keyboard': How one Ukrainian IT specialist exposed a notorious Russian ransomware gang," https://edition.cnn.com/2022/03/30/politics/ukraine-hack-russian-ransomware-gang/index.html.

[74] B. Krebs, "Pro-Ukraine 'protestware' pushes antiwar ads, geo-targeted malware," https://krebsonsecurity.com/2022/03/pro-ukraine-protestware-pushes-antiwar-ads-geo-targeted-malware/.

[75] B. Bakić, M. Milić, I. Antović, D. Savić, and T. Stojanović, "10 years since Stuxnet: What have we learned from this mysterious computer software worm?" in *IT 2021*, 2021.

[76] DoJ, "Six Russian GRU officers charged in connection with worldwide deployment of destructive malware and other disruptive actions in cyberspace," https://www.justice.gov/opa/pr/six-russian-gru-officers-charged-connection-worldwide-deployment-destructive-malware-and.

[77] J. Wolff, "Should insurance companies pay out for damage caused by state-sponsored cyberattacks?" https://slate.com/technology/2022/01/merck-notpetya-cyberattack-insurance-russia.html.

[78] "Foreign Office Minister condemns Russia for NotPetya attacks," https://www.gov.uk/government/news/foreign-office-minister-condemns-russia-for-notpetya-attacks, 2018.

[79] Security Encyclopedia, "NotPetya," https://www.hypr.com/notpetya/.

[80] C. Cimpanu, "Microsoft: Data-wiping malware disguised as ransomware targets Ukraine again," https://therecord.media/microsoft-data-wiping-malware-disguised-as-ransomware-targets-ukraine-again/.

[81] ——, "Ukraine dismantles social media bot farm spreading "panic"," https://therecord.media/ukraine-dismantles-social-media-bot-farm-spreading-panic/.

[82] "Cyber attacks on Ukraine: DDoS, new data wiper, cloned websites, and Cyclops Blink," https://behaviour-group.com/PT/cyber-attacks-on-ukraine-ddos-new-data-wiper-cloned-websites-and-cyclops-blink.

[83] J. A. Guerrero-Saade, "AcidRain — A Modem Wiper Rains Down on Europe," https://www.sentinelone.com/labs/acidrain-a-modem-wiper-rains-down-on-europe/.

[84] M. T. Klare, "Cyber battles, nuclear outcomes? Dangerous new pathways to escalation," https://www.armscontrol.org/act/2019-11/features/cyber-battles-nuclear-outcomes-dangerous-new-pathways-escalation.

[85] N. Bose, "Biden: If U.S. has 'real shooting war' it could be result of cyber attacks," https://www.reuters.com/world/biden-warns-cyber-attacks-could-lead-a-real-shooting-war-2021-07-27/.

[86] D. Stuttard and M. Pinto, *The Web Application Hacker's Handbook: Finding and Exploiting Security Flaws*, 2nd ed. Wiley, Oct 2011.

[87] P. Muncaster, "Half of government security incidents caused by missing patches," https://www.infosecurity-magazine.com/news/half-government-incidents-missing/.

[88] J. Greig, "Average time to fix critical cybersecurity vulnerabilities is 205 days: report," https://www.zdnet.com/article/average-time-to-fix-critical-cybersecurity-vulnerabilities-is-205-days-report/, 2019.

[89] F. Li and V. Paxson, "A large-scale empirical study of security patches," in *CCS 2017*.

[90] CISA, "Apache Log4j vulnerability guidance," https://www.cisa.gov/uscert/apache-log4j-vulnerability-guidance.

[91] M. Bada and I. Pete, "An exploration of the cybercrime ecosystem around Shodan," in *IOTSMS 2020*.

[92] S. Rosenblatt, "How the shady zero-day sales game is evolving," https://www.darkreading.com/edge-articles/how-the-shady-zero-day-sales-game-is-evolving, 2021.

[93] S. B. Wicker, "The ethics of zero-day exploits—: the NSA meets the trolley car," *Commun. ACM*, vol. 64, no. 1, 2020.

[94] L. H. Newman, "The leaked NSA spy tool that hacked the world," https://www.wired.com/story/eternalblue-leaked-nsa-spy-tool-hacked-world/, 2018.

[95] SentinelOne, "Eternalblue exploit: What it is and how it works," https://www.sentinelone.com/blog/eternalblue-nsa-developed-exploit-just-wont-die/, 2019.

[96] A. McNeil, "How did the WannaCry ransomworm spread?" https://blog.malwarebytes.com/cybercrime/2017/05/how-did-wannacry-ransomworm-spread/, 2019.

[97] D. Bisson, "NotPetya: Timeline of a ransomworm," https://www.tripwire.com/state-of-security/security-data-protection/cyber-security/notpetya-timeline-of-a-ransomworm/.

[98] T. Tervoort, "Zerologon:Unauthenticated domain controller compromise by subverting Netlogon cryptography (CVE-2020-1472)," https://www.secura.com/uploads/whitepapers/Zerologon.pdf, 2020.

[99] C. Bing, "China's government is keeping its security researchers from attending conferences," https://www.cyberscoop.com/pwn2own-chinese-researchers-360-technologies-trend-micro/, 2018.

[100] "Chinese government lays out new vulnerability disclosure rules," https://therecord.media/chinese-government-lays-out-new-vulnerability-disclosure-rules/.

[101] X. Shen, "Apache Log4j bug: China's industry ministry pulls support from Alibaba Cloud for not reporting flaw to government first," https://www.scmp.com/tech/big-tech/article/3160670/apache-log4j-bug-chinas-industry-ministry-pulls-support-alibaba-cloud, 2021.

[102] NIST, *Security and Privacy Controls for Information Systems and Organizations*, Sep 2020. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf

[103] "Cyber security programs for nuclear facilities," https://scp.nrc.gov/slo/regguide571.pdf, 2010.

[104] H. Assal and S. Chiasson, "Security in the software development lifecycle," in *SOUPS 2018*.

[105] P. Morrison, "A security practices evaluation framework," in *ICSE 2015*.

[106] N. Tomas, J. Li, and H. Huang, "An empirical study on culture, automation, measurement, and sharing of DevSecOps," in *Cyber Security 2019*.

[107] J. A. Morales, T. P. Scanlon, A. Volkmann, J. Yankel, and H. Yasar, "Security impacts of sub-optimal DevSecOps implementations in a highly regulated environment," in *ARES '20*.

[108] G. Wurster and P. van Oorschot, "The developer is the enemy," in *NSPW 2008*.

[109] J. Xie, H. R. Lipford, and B. Chu, "Why do programmers make security errors?" in *VL/HCC 2011*.

[110] Y. Acar, S. Fahl, and M. L. Mazurek, "You are not your developer, either: A research agenda for usable security and privacy research beyond end users," in *SecDev 2016*.

[111] M. Green and M. Smith, "Developers are not the enemy!: The need for usable security APIs," *IEEE S&P 2016*, vol. 14, no. 5.

[112] C. Weir, I. Becker, J. Noble, L. Blair, M. A. Sasse, and A. Rashid, "Interventions for long-term software security: Creating a lightweight program of assurance techniques for developers," *SPE*, vol. 50, no. 3.

[113] M. Almansoori, J. Lam, E. Fang, K. Mulligan, A. G. Soosai Raj, and R. Chatterjee, "How secure are our computer systems courses?" in *ICER 20*.

[114] C. Sadowski, E. Aftandilian, A. Eagle, L. Miller-Cushon, and C. Jaspan, "Lessons from building static analysis tools at Google," *Commun. ACM*, vol. 61, no. 4, 2018.

[115] G. P. De Francesco, "The general data protection regulation's practical impact on software architecture," *Computer*, vol. 52, no. 4, 2019.

[116] EU, "Directive (EU) 2016/1148 of the European Parliament and of the Council," https://eur-lex.europa.eu/eli/dir/2016/1148/oj.

[117] ENISA, *Advancing software security in the EU: the role of the EU cybersecurity certification framework.*, 2020.

[118] HM Government, "Government Cyber Security Strategy."

[119] A. Naiakshina, A. Danilova, E. Gerlitz, E. von Zezschwitz, and M. Smith, ""If you want, I can store the encrypted password": A password-storage field study with freelance developers," in *CHI 2019*.

[120] A. Greenberg, "The full story of the stunning RSA hack can finally be told," https://www.wired.com/story/the-full-story-of-the-stunning-rsa-hack-can-finally-be-told/.

[121] T. W. Thomas, M. Tabassum, B. Chu, and H. Lipford, "Security during application development: An application security expert perspective," in *CHI 2018*.

[122] J. Haney, M. Theofanos, Y. Acar, and S. Spickard Prettyman, ""we make it a big deal in the company": Security mindsets in organizations that develop cryptographic products," in *SOUPS 2018*.

[123] S. Rahaman, G. Wang, and D. D. Yao, "Security certification in payment card industry: Testbeds , measurements , and recommendations," *ACM CCS*, p. 481–498, 2019.

[124] L. Vaas, "BillQuick Billing App Rigged to Inflict Ransomware," https://vulners.com/threatpost/THREATPOST:94E54481AD472743701D499DC7677008.

Penny Wise, Pound Foolish: An Experimental Design of Technology Trust Amongst Organizational Users

Pratim Milton Datta
Amabassador Crawford College of Business and Economics
Kent State University, USA, University of Johannesburg, South Africa,
pdatta@kent.edu

Thomas Acton
Business Information Systems
J.E. Cairnes School of Business & Economics
NUI Galway, Galway, Ireland
thomas.acton@nuigalway.ie

Noel Carroll
Business Information Systems
J.E. Cairnes School of Business & Economics
NUI Galway, Galway, Ireland
noel.carroll@nuigalway.ie

Abstract*— *In the face of burgeoning cybersecurity and ransomware attacks, is cybersecurity technology the panacea? Building on behavioral economics, particularly moral hazard and Peltzman effects, this research uses a pilot field-experiment to investigate whether cybersecurity technology trust, in departure from general intuition, can make users "penny-wise, pound-foolish," where technology trust and our growing information needs may erode user-caution, leaving us more vulnerable.

Keywords*—*Cybersecurity, Experimental Design, Peltzman effect, Moral Hazard, Technology, Institutional Trust

I. Introduction

In an era of heightened uncertainty and urgency around disruption, digital transformation has become a global priority on leadership agendas. Digital transformation, punctuated by the COVID-19 crisis as organizations have attempted to stay operational by embracing any technology marketed as a "virtual panacea" has created a technology wild-west. But just as the pandemic accelerated the need for change through digital transformation, it has exposed serious cybersecurity vulnerabilities [1, 2, 16].

While organizations and companies have rapidly embraced digital transformation during the COVID-19 crisis, society has seen a 600%+ increase in malicious emails. With employees, clients, and vendors leapfrogging to virtual meetings and work-from-home access and BYOD (Bring You Own Device), the pandemic has witnessed a tenfold increase in ransomware (e.g. UCSF's medical school) and a 200%+ increase in Denial of Service (DoS) attacks (e.g. a 2.3 terabyte attack on AWS). In fact, MIT estimates global cybercrime costs at US$5 Trillion, stemming from information-stealing scams, ransomware, and work-from-home vulnerabilities [1, 2].

However, cyberattacks during the pandemic were not for want of cybersecurity investments. In the same period, cybersecurity technology spending increased by nearly 38 percent from 40.8 in 2019 to $55 billion [1, 3]. Cybersecurity technology startup funding have also increased from $6.9 billion in 2020 to $17.4 billion in venture dollars in 2021 [3]. Despite increased spending and investments in the cybersecurity technology space, cyberattacks have dramatically increased.

The paradox of increasing cybercrime in the face of increasing cybersecurity technology spending begs investigation. Motivated by this paradox, our research investigates the paradoxical effects of cybersecurity technology deployments on cybersecurity user behaviors in organizations.

II. Theoretical Underpinnings and Hypothesis Development

A. Moral Hazard

Moral hazard is an economic principle that aims to explain why individuals engage in high levels of risk-taking behavior when they feel that the consequences of such behaviors are minimal. Popularized by Arrow [5], the moral hazard principle has been used to examine various risk-taking and adverse selection behaviors across a multitude of phenomena [5, 6, 7]. For example, users that purchase rental car insurance or health insurance may feel protected from consequences, thus prompting users to assume "morally hazardous" choices by greater risks than they would have otherwise undertaken.

From a cybersecurity standpoint, the continuous global use of unsupported Windows OS instances such as WinXP and Windows Vista or the use of Adobe Flash open a plethora of system vulnerabilities. Yet, users continue to adversely select and use such instances owing to certain conveniences (e.g. the unwillingness to purchase and install newer versions, the ability to run legacy applications, among others).

B. Peltzman Effect and User Caution

In 1975, Peltzman [8] conducted a longitudinal time-series study using the US' NHTSA (National Highway Transportation and Safety Administration) data to offer an insight that became known as the Peltzman effect. The study found that automobile technological advancements (seat belts, airbags), intended to reduce automobile accidents, unintendedly created an unexpected moral hazard. Drivers, feeling safe from such technological advancements, often drove complacently, resulting in an increase in minor accidents and pedestrian injuries.

The Peltzman effect, as an instantiation of cybersecurity moral hazard [10, 11], may help explain why several recent cybersecurity breaches have occurred in organizations that themselves offer cybersecurity and infrastructure management solutions such as the 2020 Solarwinds' Orion server hack [11]. As Datta [11] remarks, "user (including vendor and consumer) errors are the weakest links, regardless of whether the user error

is analog or automated (embedded in the operational logic)." So, it stands to reason why cybersecurity threats often originate "with phishing and spoofing attacks intended to exploit and manipulate human psychology rather than technology" [10].

C. Cybersecurity Technology Trust and User Caution

User trust in cybersecurity technologies is salient to our investigation of user behavior instantiations in light of moral hazards and the Pelzman effect. Cybersecurity technologies offer an institutional aegis under which users operate. Cybersecurity technologies are not a single artifact (similar to dyadic trust) but a portfolio of technologies (intrusion detection, spam filters, phishing detectors, Firewalls, DMZ, honeytokens, etc., along with internal policies, guidelines, and operations) that offer an institutional context. Deductively, institutional trust offers the appropriate trust lens to examine cybersecurity user-behavior.

Based on work by [12, 13], institutional trust in cybersecurity technologies is the willingness of individuals to be vulnerable to overarching cybersecurity technological operations, rules, and regulations. Inasmuch as organizations use cybersecurity technologies to offer an institutional aegis, institutional trust triggers a trust calculus that transfers trust to cybersecurity technologies. We posit that such trust transference begets a sense on technological overreliance and a sense of laxity and complacency that hackers and other cybercriminals are quick to capitalize. Such was the case in the global Solarwinds' Orion Server hack where a misplaced password and a complacent GitHub post allowed hackers to inject malware [10, 11].

In summary, we argue that users make sub-optimal choices from complacency when perceiving that cybersecurity technologies offer a panacea and will protect them from any cybersecurity threats. Thus, we postulate:

H1: There is in inverse relationship between cybersecurity technology trust and user-caution related to cybersecurity such that users who perceive a higher level of trust in cybersecurity technology protection will demonstrate a lower level of user-caution.

D. Information Specificity

Rooted in Williamson's [14] notion of asset specificity that access to information to perform a task or transaction is a key asset in any exchange, use information specificity as "the demand and reliance placed by a consumer on information that is specific and timely to assist them in their tasks and transactions" [13, 18].

Information access has become the *sine qua non* of our everyday existence. Users place higher information specificity on information that are core to their decision-making and operations and low specificity on information that are peripheral. For example, an organizational user such as an account manager may place higher information specificity on billing and A/R information and lower information specificity on wellness initiative information.

In 2014, around the time the Russian troops were moving into Crimea, a malware injection was triggered by an injudicious download of a program assumed to be a legitimate artillery software update that would have been critical to Ukrainian military operations. Likewise, recent ransomware attacks where hackers encrypt vital (high specificity) information for ransom (e.g the Colonial Pipeline attack or Popp's AIDS disk [10, 11]) exemplify how user behaviors are contingent upon asset specificity.

Further, user behavior can become more chaotic in instances where the user cannot obtain important information in a timely fashion, at times resorting to less-secure actions to do so, for example, asking a colleague to email a database to them because they cannot log in directly. Patently, we reason that, *ceteris paribus,* denying users access to a certain piece of information with high information specificity may trigger hastier behaviors with less caution because of a frantic need for information access.

Regardless of a user's institutional trust in cybersecurity technologies, higher information specificity can prompt frantic and injudicious user behavior where caution is foregone for information access. Thus, we advance:

H2: User information specificity moderates the relationship between Cybersecurity Technology Trust and User-Caution. The moderation is such that, given the same level of Cybersecurity Technology Trust, users with higher information specificity will exhibit lower user-caution.

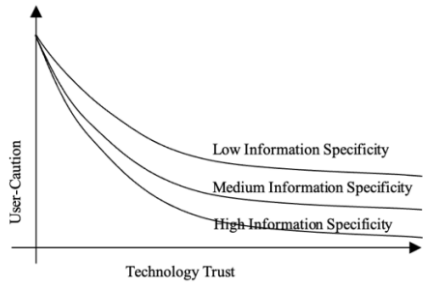


Figure 1: The Hypothesized Moderating Influence of Information Specificity on Cybersecurity Technology Trust and User-Caution.

III. Experimental Design and Results

The study uses a field experiment. The field experiment was administered in a multinational manufacturing organization in the Northeast US between November and December of 2021. The organization allowed us to pilot the experiment as a part of a larger study to assess cybersecurity user-behavior.

In our field-experiment design, we controlled for user self-efficacy, given the moderating influence and endogeneity surrounding user self-efficacy. Based on the work by [4], self-efficacy is the users' belief in their competence related to understanding and using a specific artifact. Given that cybersecurity is rapidly evolving topic, it was important to control for cybersecurity self-efficacy. Thus, the study was particularly cautious in ensuring that the field-experiment participants had a minimum level of cybersecurity self-efficacy.

All pilot study participants belonged to the IT support or Data Center groups. To ensure a minimum-acceptable-level of cybersecurity efficacy, i.e., individual user belief about one's cybersecurity competence, all participants were required to complete a 10-question proprietary cybersecurity competence evaluation questionnaire, enabling the calculation of a cybersecurity efficacy score. Of the 81 participants, 43 users were selected for the pilot based on a cybersecurity efficacy score >75%. For the pilot, we wanted to validate that all

participants (n=43) were cyber-literate to eliminate variances within a small sample. This validity-check reduced spurious effects and allowed us to control for efficacy on user-behavior. Our high-efficacy requirement also helped our instrument development by drawing upon expert feedback for content and construct validity. We plan to remove efficacy checks for future large-scale data collection and use efficacy as a control variable.

Participants were 42% female (18), with 45.6 years average age and 12.1 years' average experience. 68% (29) of the participants belonged to the Data Center operations group.

A. Experimental Dashboard Design

Figure 2a: Experiment Dashboard screenshot where Company Cybersecurity Technology Threat Mitigation Performance ≅ Industry Cybersecurity Technology Performance

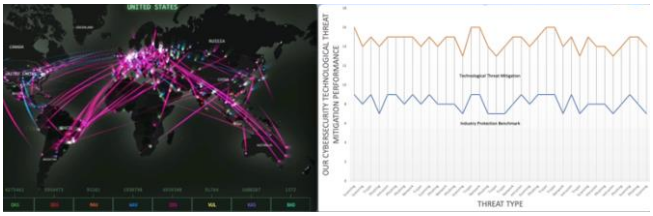


Figure 2b: Experiment Dashboard screenshot where Company Cybersecurity Technology Threat Mitigation Performance > Industry Cybersecurity Technology Performance

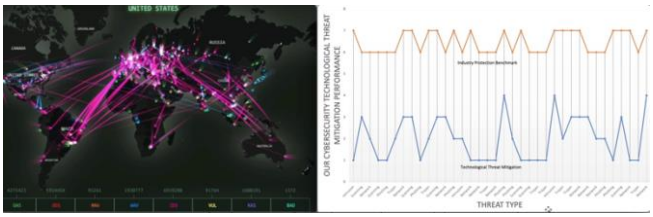


Figure 2c: Experiment Dashboard screenshot where Company Cybersecurity Technology Threat Mitigation Performance < Industry Cybersecurity Technology Performance

The field experiment design used a cybersecurity technology dashboard with three manipulations – low cybersecurity technology protection, medium cybersecurity technology protection, and high cybersecurity technology protection, all relative to industry benchmarks (see fig 2a-2c). The dashboard was a dynamic, looping video with 2 frames and 3 elements:

The left dashboard frame captured real-time cyberattacks. The right frame included 2 elements: (a) The Company's Cybersecurity Technology Threat Mitigation Performance, and (b) Industry Threat Mitigation Benchmark Performance.

We used a repeated measure experimental design where users were treated with the same variables under different conditions. A repeated measure experimental design was applicable for our empirical study because (i) it is particularly efficient with a limited number of participants because each participant is exposed to the treatment multiple times (n x 3), (ii) insofar as cybersecurity behavior is concerned, a repeated measure design affords the ability to observe users reacting to various treatments, akin to real-time workplace exposure. Finally, a repeated measure design collects longitudinal data that can be used for further, granular analysis. All 43 users were exposed to the dashboard at 3 different times (all between 10-11am) over a week with a different manipulation used for each treatment. Treatments were assigned randomly. Figures 2a-c depict the dashboards as static images.

Each dashboard was presented atop a survey followed by 7 emails as a part of the questionnaire (to confirm reading and not being lost in an inbox deluge). All emails carried spoofed organizational domains for legitimacy. With a very large PMO (Project Management Office) and hundreds of vendors, choosing these two allowing for anonymous spoofing. Each email appeared to originate from either (i) the internal organization PMO or (ii) external preferred vendors.

Email headings and content included Call Center Database Migration, Incidence Response Plan for Cybersecurity Breaches, Vendor Authentication Protocols, Business Intelligence Requirements, PMO Meeting Notes and Minutes, among others.

The 7 emails were drawn from a pool of 29 curated emails. The curation used historically authentic communication to ensure that the structure and wording followed established organizational communication standards. Curating based historically authentic communication also ensured that our emails, if relayed in real-time via an SMTP/POP server, would not be marked as junk or spam based on technology filters.

Each email included either a link or a document. A 9-point Likert scale was used to measure user caution, with 1-3 representing users' willingness to click/download (absence of caution to low user-caution), 4-6 for "disregard with no inquiry," "disregard and perhaps inquire," and "disregard and inquire immediately" (low to medium user-caution), and 7-9 for "flag and reach" (medium to high user-caution).

As the onset of the experiment, users were notified the fact that all 7 emails were "important" communications and each needed responding. 4 of the 7 emails (57%) were fraudulent.

Informed consent was used. Users were made aware the need to remain vigilant in face of deceptive communications (e.g. phishing, malware). Users had the ability to hover over links or documents to assess safety. Fraudulent links had obvious paths (e.g., *s30972.co.ru/PMO*) and fraudulent documents were attached files (e.g. .doc files) with ambiguous names (e.g. *"project_rewards"*). Scales for Information Specificity were drawn from pre-validated measures by [15].

B. Exploratory Results

The repeated measure design yielded *n* x *3* (43 x 3 = 129) observations. An initial ANOVA between-treatment analysis demonstrated no statistically significant differences ($p>0.05$) between users exposed to the high, medium, and low cybersecurity technology treatments. Since no mean differences were found, we averaged user caution values for each treatment. While a lack of differences between means reduced the sample size back to n=43, it confirmed consistent user-behavior between treatments.

Participants' Technology Trust ($\bar{x}$=4.65, σ=1.79) had relatively higher trust distribution (-ve skewness); User

Caution ($\bar{x}$=3.89, σ=1.23) had relatively lower caution distribution (+ve skewness), and Information Specificity ($\bar{x}$=4.47, σ=2.01) had relatively higher specificity (-ve skewness).

TABLE I. CONSTRUCT STATISTICS AND CORRELATIONS

Construct Correlations	Information *Specificity*	*Technology Trust*	*User Caution*
Information Specificity	1		
Technology Trust	0.764985261	1	
User Caution	-0.831508304	-0.862164871	1

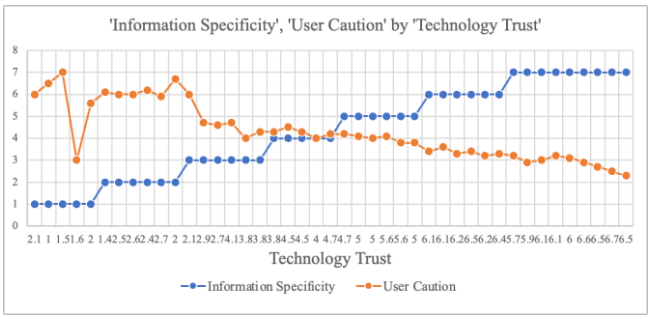


Figure 3a: Scatter Chart of the Interplay of Cybersecurity Technology Trust, User-Caution, and Information Specificity

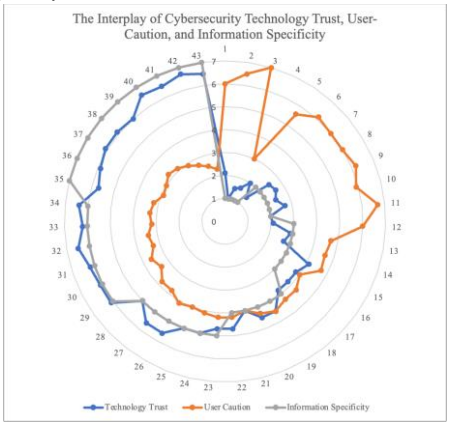


Figure 3b: Radar Charts of the Interplay of Cybersecurity Technology Trust, User-Caution, and Information Specificity

The correlation matrix (Table 1) along with the radar chart (Fig 3) from the pilot study offer initial insights on the interplay of Cybersecurity Technology Trust (CTT), User Caution (UC), and Information Specificity (IS). Essentially, as hypothesized, user caution is inversely correlated with cybersecurity technology trust, further accentuated in the presence of information specificity.

Given the small pilot sample size (this lacking power and increasing Type II errors), it would be erroneous to draw robust inferences. However, for exploratory purposes, both H1 and H2 are supported in terms of hypothesized directionality but not significance. Exploratory results show CTT—UC (-0.359, ns); TS—UC (-0.071, ns); and TS*CTT—UC (-0.021, ns).

To conclude, in line with moral hazard and Peltzman effects, over-reliance on cybersecurity technology may be akin to being penny-wise, pound-foolish. A perceived overreliance on cybersecurity technologies may prompt an unconscious user-caution, especially in the face of growing information-specificity. Inadvertently, too much cybersecurity technology reliance may expose an organization to greater user vulnerabilities.

IV. FUTURE DIRECTIONS

This paper outlined first steps in the relationship and interplay between cybersecurity trust, information specificity and user-caution. We described a preliminary experiment to assess two initially-supported hypotheses. Next steps are to expand the data set, refine the experimental instrumentation and analysis structure, and provide expanded statistical evidence for the hypotheses.

Our finds find that user laxity in the face of cybersecurity technology reliance is anything but exculpatory. For academia and practice, our findings question whether organizational overemphasis on their cybersecurity investments can inadvertently induce user prodigality, leaving organizations "penny wide, pound foolish!"

REFERENCES

[1] PurpleSec "Cyber Security Statistics: The Ultimate List Of Stats, Data & Trends",, 2021 url: https://purplesec.us/resources/cyber-security-statistics/

[2] S. Brown, "How to think about cybersecurity in the era of COVID-19." MIS Sloan Ideas, Aug. 20. 2020, url: https://mitsloan.mit.edu/ideas-made-to-matter/how-to-think-about-cybersecurity-era-covid-19. .

[3] Statista, Spending on cybersecurity worldwide from 2017 to 2021 (COVID-19 adjusted). url: https://www.statista.com/statistics/991304/worldwide-cybersecurity-spending/

[4] D. R. Compeau, and C. A. Higgins. "Computer Self-Efficacy: Development of a Measure and Initial Test." MIS Quarterly, vol. 19, no. 2, Management Information Systems Research Center, University of Minnesota, 1995, pp. 189–211, https://doi.org/10.2307/249688.

[5] K. Arrow. "Uncertainty and the Welfare Economics of Medical Care". The American Economic Review. American Economic Association. 53 (5):1963, pp. 941–73.

[6] M.V. Pauly. "The economics of moral hazard: comment". The American Economic Review. American Economic Association. 58 (3): 1968, pp. 531–37.

[7] D Rowell, L.B. Connelly, "A history of the term 'moral hazard'" Journal of Risk and Insurance 79 (4), 2012, pp. 1051–75Y.

[8] S. Peltzman, "The Effects of Automobile Safety Regulation." *Journal of Political Economy*, vol. 83, no. 4, University of Chicago Press, 1975, pp. 677–725,

[9] S. Bakshi. "The Peltzman Effect and Cybersecurity" *ISACA Blog*, 2021. url: https://www.isaca.org/resources/news-and-trends/newsletters/atisaca/2021/volume-16/the-peltzman-effect-and-cybersecurity

[10] P. Datta. "Cyberruse at the Cybergates Technology, People and Processes" *ISACA journal*, Vol. 6, 2021, pp. 51-56.

[11] P. Datta. "Hannibal at the gates: Cyberwarfare & the Solarwinds sunburst hack." *Journal of Information Technology Teaching Cases*. March 2021. doi:10.1177/2043886921993126.

[12] D.H.McKnight, L.L. Cummings, and N.L. Chervany, "Initial trust formation in new organizational relationships." *Academy of Management Review* 23(3), 1998, pp. 473–490.

[13] P. Datta and S. Chatterjee. "The economics and psychology of consumer trust in intermediaries in electronic markets: the EM-Trust Framework." European Journal of Information Systems 17, no. 1 (2008): 12-28.

[14] O.E. Williamson. "The economics of organization: The transaction cost approach." *The American Journal of Sociology* 87(3), 1981, pp. 548–577.

[15] P. Datta and S. Chatterjee. "Online consumer market inefficiencies and intermediation." *SIGMIS Database*, 42, 2 (May 2011), pp. 55–75.

[16] L. M. Giermindl, F. Strich, O. Christ, U. Leicht-Deobald, and A. Redzepi, "The dark sides of people analytics: reviewing the perils for organisations and employees," *European Journal of Information Systems,* pp. 1-26, 2021, doi: 10.1080/0960085X.2021.1927213

Gradient Information From Google GBoard NWP LSTM Is Sufficient to Reconstruct Words Typed

Mohamed Suliman, Douglas J. Leith
Trinity College Dublin, Ireland

***Abstract*—Federated Learning is now widely deployed by Google on Android handsets for distributed training of neural networks. While Federated Learning aims to avoid sharing sensitive user data with Google, in this paper we show that when used for GBoard next word prediction Federated Learning provides little privacy to users. Namely, we demonstrate that the words typed by a user can be quickly and accurately reconstructed from the gradients of the GBoard LSTM used for next word prediction. Use of mini-batches does not protect against reconstruction.**

I. Introduction

In [1] Google introduced Federated Learning for privacy-enhanced distributed training of neural networks. Federated Learning is now widely deployed on Android mobile handsets, and in particular is used for next word prediction in Google's GBoard keyboard app [2] which, according to the Google Play store, is installed in more than 1 Billion devices[1]. In Federated Learning a central server collects gradient vectors from mobile handsets running the model, it executes a stochastic gradient descent step to update the model parameters and then pushes the new parameter values to the mobile handsets. This process repeats until the parameters are judged to have converged, a specified number of iterations have been completed etc. By keeping the training data on the mobile handsets and only sharing gradient information, the hope is that a degree of privacy is gained. However, there has been little formal privacy analysis of Federated Learning.

In this paper we show that when used for GBoard next word prediction Federated Learning provides little privacy to users. Namely, we demonstrate that the words typed by a user can be quickly and accurately reconstructed from the gradients of the GBoard LSTM used for next word prediction. Use of mini-batches does not protect against reconstruction.

Key to the lack of privacy is that in next word prediction the neural net input is echoed by the neural net output. That is, in the next word prediction task the output of the neural net aims to match the sequence of words typed by the user, albeit with a shift one word ahead. The sign of the output loss gradient directly reveals information about the words typed by the user, which can then be reconstructed by inspection (there is no need for any complex processing).

We provide more details below, but briefly illustrate the nature of the information leakage here. Let D be a dictionary of V words, each typed word is mapped to an entry in D and the next word prediction is a vector $\hat{y} \in [0,1]^V$ whose i'th element $\hat{y}_i$ is the probability that the next word will be the i'th dictionary entry. When a softmax output layer is used $\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{V} e^{z_j}}$ with vector $z \in \mathbb{R}^V$ the raw logits. Suppose the next word typed by the user is entry i^* from the dictionary. The cross-entropy loss is $J = -\log \frac{e^{z_{i^*}}}{\sum_{j=1}^{V} e^{z_j}}$. The derivative $\frac{\partial J}{\partial z_{i^*}}$ is $\frac{e^{z_{i^*}}}{\sum_{j=1}^{V} e^{z_j}} - 1 < 0$ while for $i \neq i^*$ the derivative is $\frac{\partial J}{\partial z_i} = \frac{e^{z_i}}{\sum_{j=1}^{V} e^{z_j}} > 0$. Hence, we can infer the index i^* of the word typed by the user simply by inspecting the sign of the loss derivatives. This is illustrated schematically in Figure 1. Federated Learning shares the derivatives of the loss with respect to model parameters, rather than the derivatives with respect to the logits y_i, but the derivatives of parameters in the penultimate output layer are enough.

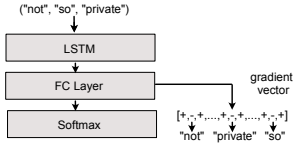


Fig. 1. Illustrating how sign of gradient elements can leak words typed.

Previous work on information leakage by Federated Learning has mainly focused on object detection tasks, e.g. see [3], [4], [5]. The neural network output is an object label (e.g. cat, dog) and the attack aims to reconstruct the input image only from gradient information. The attack is typically formulated as an optimisation problem and solved using gradient descent. Successful attacks have been demonstrated for standard neural nets and image datasets. However, because in next word prediction the neural net input is echoed by the output we are able to perform much faster, more robust attacks that are difficult to defend against. We demonstrate the effectiveness of these attacks against a widely deployed LSTM neural net from GBoard.

II. Privacy Threat Model

The transmission of user data from mobile handsets to back-end servers is not intrinsically a breach of privacy. For instance, it can be useful to share details of the device model/version and the locale/country of the device when checking for software updates. This poses few privacy risks if the data is common to many handsets and therefore cannot be easily linked back to a specific handset/person [6], [7].

Two major issues in handset privacy are (i) release of sensitive data, and (ii) de-anonymisation i.e. linking of data to a person's real world identity.

Release of sensitive data. What counts as sensitive data is a moving target, but it seems clear that the words entered by users, e.g. when typing messages, writing notes and emails, web browsing and performing searches, may well be sensitive. It is not just the sentences typed which can be sensitive but also just the list of words used (i.e. even without knowing the word ordering) since this can be used for targeting surveillance via keyword blacklists [8]. It is also important to note that data which is not sensitive in isolation can become sensitive when combined with other data, and this is a particular concern with regard to large companies that operate mobile payment services, supply web browsers, run advertising platforms etc.

De-anonymisation. Android handsets can be directly tied to a person's real identity in several ways, even when a user

This work was supported by SFI grant 16/IA/4610.

[1]https://play.google.com/store/apps/details?id=com.google.android.inputmethod.latin, accessed 17th Feb 2022.

takes active steps to try to preserve their privacy. Probably most relevant here is via the Android ID, since most Google telemetry is tagged with this. Via other data collected by Google Play Services the Android ID is linked to (i) the handset hardware serial number, (ii) the SIM IMEI (which uniquely identifies the SIM slot) and (iii) the user's Google account [9], [10]. When creating a Google account it is necessary to supply a phone number on which a verification text can be received. For many people this will be their own phone number. Use of Google services such as buying a paid app on the Google Play store or using Google Pay further links a person's Google account to their credit card/bank details. A user's Google account, and so the Android ID, can therefore commonly be expected to be linked to the person's real identity.

III. Related Work

The use of Federated Learning for next word prediction in Google's GBoard app is considered in [2] and the structure of the LSTM that we extracted from the app matches the description in that paper. The Long Short Term Memory (LSTM) recurrent neural network was introduced in [11], and a variant termed the Coupled Input Forget Gate (CIFG) LSTM was developed in [12]. CIFGs include less parameters that need to be trained than a regular LSTM, and are thus appealing for mobile handset deployment, where network bandwidth and storage resources cannot be guaranteed.

Since being introduced by [1], Federated Learning has attracted a great deal of interest and generated a growing body of literature, in particular about the security challenges it poses [13], [14], [15]. The attack presented herein can be classed as an *inference attack*, where training inputs and labels are inferred from the model updates. Information leakage from the gradients of neural nets used for object detection appears to have been initially investigated in [3], which proposed a so-called Deep Leakage from Gradients (DLG) method for input image reconstruction. This work was subsequently extended by [5], [4]. In particular, [4] demonstrated effective input image reconstruction even for mini-batch sizes up to 48 images. Other inference attacks attempt to infer a particular data point's membership in the training data [16], [17], [18], as well properties of other participants' training data [17].

IV. GBoard Next Word Prediction LSTM

A. LSTM Software Version & Tensorflow Implementation

We used a rooted Google Pixel 2 running Android 11 and Google GBoard app version 10.5.03.367007960. We extracted the files nwp.csym, nwp.csym2 and nwp.uint8.mmap.tflite from folder files/superpacks/next-word-predictor/tflite-nwp-45c6579035e9df4ffe5c1246f8d5615d within the app private data directory. The file nwp.csym2 contains the LSTM dictionary containing $V = 9502$ words. The dictionary is stored in MARISA-Trie format[2] with an additional pre-pended file header. The file nwp.uint8.mmap.tflite is the LSTM, stored in TensorFlow Lite format[3]. The publicly available version of TensorFlow Lite does not support training or the calculation of model derivatives. We therefore ported the TensorFlow Lite model to TensorFlow, verifying that both generated identical outputs for the same inputs. The weights in the TensorFlow Lite model are tagged with descriptive labels, and are used to create the replicated model in TensorFlow.

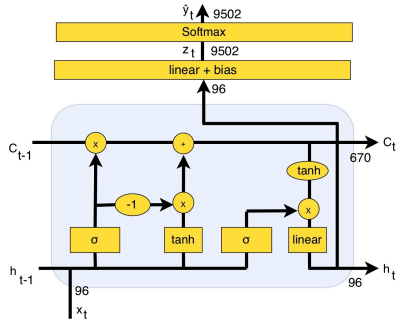


Fig. 2. Schematic of LSTM architecture. LSTM layer takes as input dense vector x_t representing a typed word and outputs a dense vector h_t. This output is then mapped to vector z_t of size 9502 (the size of the dictionary) with the value of each element being the raw logit for the corresponding dictionary word. A softmax layer then normalises the raw z_t values to give a vector $\hat{y}_t$ of probabilities.

B. LSTM Architecture

Input words are first mapped to a dictionary entry, with a special `<UNK>` entry used for words that are not in the dictionary. The index of the dictionary entry is then mapped to a dense vector of size $D = 96$ using a lookup table (the dictionary entry is one-hot encoded and then multiplied by an $\mathbb{R}^{D\times V}$ weighting matrix W^T) and applied as input to an LSTM layer with 670 units i.e. the state C_t is a vector of size 670. The LSTM layer uses a CIFG architecture without peephole connections, illustrated schematically in Figure 2. The LSTM state C_t is linearly projected down to an output vector h_t of size D, which is mapped to a raw logit vector z_t of size V via a weighting matrix W and bias b. This extra linear projection is not part of the orthodox CIFG cell structure, and is included to accommodate the model's tied input and output embedding matrices [19]. A softmax output layer finally maps this to an $[0, 1]^V$ vector $\hat{y}_t$ of probabilities, the i'th element being the estimated probability that the next word is the i'th dictionary entry.

C. Loss Function

Following [2] we use categorical cross entropy loss over the output and target labels.

V. The Attack: Gradient Information Leakage

A. Recovering The Words Typed

After the user has typed t words the output of the neural net is next word prediction vector $\hat{y}_t$,

$$\hat{y}_{t,i} = \frac{e^{z_i}}{\sum_{j=1}^{V} e^{z_j}}, \ i = 1, \ldots, V$$

with raw logit vector $z_t = Wh_t + b$, where h_t is the output of the LSTM layer. The cross-entropy loss function for text consisting of T words is $J_{1:T}(\theta) = \sum_{t=1}^{T} J_t(\theta)$ where

$$J_t(\theta) = -\log \frac{e^{z_{i^*}(\theta)}}{\sum_{j=1}^{V} e^{z_j(\theta)}}$$

where i_t^* is the dictionary index of the t'th word entered by the user and θ is the vector of neural net parameters (including

[2]See, for example, https://android.googlesource.com/platform/external/marisa-trie/

[3]https://www.tensorflow.org/lite/

the elements of W and b). Differentiating with respect to the output bias parameters b we have that,

$$\frac{\partial J_{1:T}}{\partial b_k} = \sum_{t=1}^{T}\sum_{i=1}^{V} \frac{\partial J_t}{\partial z_{t,i}} \frac{\partial z_{t,i}}{\partial b_k}$$

where

$$\frac{\partial J_t}{\partial z_{t,i_t^*}} = \frac{e^{z_{i^*}}}{\sum_{j=1}^{V} e^{z_j}} - 1 < 0 \quad \frac{\partial J_t}{\partial z_{t,i}} = \frac{e^{z_i}}{\sum_{j=1}^{V} e^{z_j}} > 0, \ i \neq i_t^*$$

and

$$\frac{\partial z_{t,i}}{\partial b_k} = \begin{cases} 1 & k = i \\ 0 & \text{otherwise} \end{cases}$$

That is,

$$\frac{\partial J_{1:T}}{\partial b_k} = \sum_{t=1}^{T} \frac{\partial J_t}{\partial z_{t,k}}$$

It follows that for words k which do not appear in the text $\frac{\partial J_{1:T}}{\partial b_k} > 0$. Also, assuming that the neural net has been trained to have reasonable performance then e^{z_k} will tend to be small for words k that do not appear next and large for words which do. Therefore for words i^* that appear in the text we expect that $\frac{\partial J_{1:T}}{\partial b_{i^*}} < 0$.

Example: Suppose the input to the neural net is the sentence "this online learning is not so private". Calculating the gradients $\frac{\partial J_{1:T}}{\partial b_k}$, $k = 1, \ldots, V$, sorting the values in descending order and selecting the elements with negative values yields the following (k'th word, $\frac{\partial J_{1:T}}{\partial b_k}$) pairs: ("learning",-0.9997006) ("private",-0.9994907) ("online",-0.99600935) ("not",-0.9627077) ("so",-0.94802666) ("is",-0.78382504). All other words in the dictionary have non-negative gradients.

This observation is intuitive from a loss function minimisation perspective. Typically the estimated probability $\hat{y}_{i^*}$ for an input word will be less than 1. Increasing $\hat{y}_{i^*}$ will therefore decrease the loss function i.e. the gradient is negative. Conversely, the estimated probability $\hat{y}_i$ for a word that does not appear in the input will be small but greater than 0. Decreasing $\hat{y}_i$ will therefore decrease the loss function i.e. the gradient is positive.

While we focus on the bias parameters b here since they yield particularly simple expressions, similar analysis applies to the W parameters and can also be expected to apply to other forms of penultimate output layer. The key is that because the output $\hat{y}_t$ aims to echo the words typed by the user, the gradient of the loss with respect to parameters in the penultimate layer will always tend to directly reveal information about the words typed (unlike in the case of object detection where the neural net output is just the object label and so reconstruction of the full input image is an additional, challenging, step).

B. Recovering The Sentences Typed

The approach above extracts the set of words typed by inspection, however this gives no indication of the original ordering of words so as to reconstruct the original sentences typed. For mini batches consisting of one sample, and short sentences, a brute force method is sufficient to reconstruct the original sentence.

Given that we have extracted n tokens, we rank all $n!$ permutations of these tokens based off of their *gradient loss*. This is defined as the L2 norm between the original FL gradient, and the gradient generated when training the model with the sentence represented by the current permutation. This loss function is used in [3] to guide the gradient descent optimization as part of the Deep Leakage from Gradients algorithm.

```
Oh k...i'm watching here:)

<S> oh k im watching here
```

Fig. 3. Here we give a sample reconstruction. The first line gives the original sentence, and the second line gives the attempted reconstruction, after extracting the tokens from the gradient information and finding the best permutation of these tokens via brute force. Note the start of sentence token <S> present in the reconstruction.

Figure 3 gives an example of the kind of reconstruction that is possible with this attack. With a mini-batch size of one, and short sentences of words contained in the model's vocabulary, reconstruction is almost perfect, albeit missing punctuation. This approach does not scale well for larger mini-batch sizes or longer sentences as the number of possible number of permutations increases. Further research is needed to look into more efficient sentence reconstruction.

VI. Performance Evaluation

A. Datasets Used

To evaluate the effectiveness of our attacks we use two datasets: (i) the UMass Global English on Twitter Dataset which contains 10,502 tweets, randomly sampled from publicly available geotagged Twitter messages [20] and (ii) a corpus of 63,632 non-Spam SMS messages [21].

B. Performance Metrics

To evaluate performance we use two metrics. Firstly, the proportion of words from the original text that the attack described in Section V-A manages to correctly reconstruct. Secondly, a modified version of the Levenshtein ratio i.e. the normalised Levenshtein distance [22] (the minimum number of word level edits needed to make one string match another) between the original text and the sentences reconstructed using the attack in Section V-B. A Levenshtein ratio closer to 100 indicates a greater match between the original and reconstructed sentences. For example, given an original sentence of "hello how are you", the reconstructions "how hello are you" and "hello how you are" both have a Levenshtein ratio of 76, as they are off by one word.

C. Mini-Batches

We evaluate performance for a range of mini-batch sizes from 1 up to 48. A mini-batch of size n consists of n separate messages from the selected dataset. At the start of each separate message the LSTM is initialised, the words for the message are input and the next word predictions noted. The sum-gradient over the n messages in a mini-batch is then used for our reconstruction attack. We consider both situations where (i) all of the messages in a mini-batch have the same number of words and (ii) where the messages may have different numbers of words in which case shorter messages are padded with the <UNK> token to match the length of the longest message in the batch (this is necessary to ensure that the gradient vectors are the same size and so can be summed).

D. Measurements

Table I shows the measured accuracy at reconstructing the words contained in a mini-batch of messages vs the number of messages in the mini-batch i.e. the mini-batch size. Note that since the input is echoed by the model's output, but shifted one word ahead, the first word is not recoverable via these means. However, since the first word is always the start of sentence

TABLE I
PROPORTION OF WORDS CORRECTLY RECONSTRUCTED.

Twitter		SMS	
Messages with 4 words			
Mini-Batch Size (#batches)	Accuracy	Mini-Batch Size (#batches)	Accuracy
1 (249 batches)	0.947	1 (155 batches)	0.985
4 (62 batches)	0.975	4 (38 batches)	0.976
8 (25 batches)	0.977	8 (17 batches)	0.983
16 (15 batches)	0.965	16 (9 batches)	0.957
32 (7 batches)	0.936	32 (4 batches)	0.933
48 (5 batches)	0.907	48 (3 batches)	0.918
Messages with 8 words			
1 (405 batches)	0.913	1 (335 batches)	0.977
4 (101 batches)	0.966	4 (83 batches)	0.965
8 (50 batches)	0.961	8 (41 batches)	0.948
16 (24 batches)	0.933	16 (19 batches)	0.926
32 (12 batches)	0.908	32 (10 batches)	0.875
48 (8 batches)	0.893	48 (6 batches)	0.858
Twitter Messages with 10 or more words			
1 (2724 batches)		0.935	
4 (681 batches)		0.961	
8 (340 batches)		0.938	
16 (170 batches)		0.917	
32 (85 batches)		0.893	
48 (56 batches)		0.885	

TABLE II
4 WORD SMS AND TWITTER SENTENCE ORDERING RESULTS WITH BATCH SIZE 1.

Dataset (batch size, #batches)	Levenshtein ratio
SMS (1, #155)	97.161 (78.7% perfect)
Twitter (1, #249)	90.173 (54.2% perfect)

token <S>, this fact is inconsequential, as we are only unable to retrieve <S>, which we assume to be included as part of a training sentence. Observe also that the accuracy remains the same as the mini-batch size and messages length increase, highlighting that use of mini-batches is an ineffective defence against this attack.

To boost performance, sentences are passed through a spell checker to find mispelled words that if spelled correctly, would represent a word that is part of the model's vocabulary. This allows the attack to reconstruct the word that was typed (albeit incorrectly), instead of just the unknown token <UNK>.

Table II reports the proportion of sentences reconstructed perfectly[4] by the brute force attack described in Section V-B. We provide results for a mini-batch size of 1 and for 4 word messages from both datasets. The Levenshtein ration measures message similarity (with 100 being considered a perfect match). SMS message reconstruction reports an average Levenshtein ration of 97.161, over 155 different messages, with 78.7% of them being reconstructed perfectly, while most other being off by at most 1 word placement. Twitter message reconstruction is slightly lower, however this is due to the high number of repetition of the unknown character, <UNK>. This corresponds to the fact that tweets often consist of unique usernames, links, hashtags, emojis, etc.

VII. SUMMARY AND CONCLUSIONS

We show that when used for GBoard next word prediction Federated Learning provides little privacy and that the words typed by a user can be quickly and accurately reconstructed. The attack itself appears to be difficult to defend against. Use of mini-batches (i.e. combining multiple messages) is demonstrated to be ineffective. Sampling the gradient vector and sending only a subset of elements is unlikely to be effective due to the large size of the gradient vector relative to the average message size i.e. to provide a reasonable defence the sampling fraction would have to be so low as to disrupt neural network training. Adding noise to the gradients is also likely to be problematic since our attack just relies on sign information and noise that disrupts the attack is also likely to disrupt neural network training. Combining gradients from multiple handsets has been proposed to improve privacy but requires co-ordination between handsets which can be difficult to achieve in practice.

[4]Excluding the first word, the start of sentence token <S>.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc AISTATS 2017*, vol. 54. PMLR, 2017, pp. 1273–1282. [Online]. Available: http://proceedings.mlr.press/v54/mcmahan17a.html

[2] A. Hard, K. Rao, R. Mathews, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated learning for mobile keyboard prediction," *CoRR*, vol. abs/1811.03604, 2018. [Online]. Available: http://arxiv.org/abs/1811.03604

[3] L. Zhu, Z. Liu, , and S. Han, "Deep leakage from gradients," in *Proc NeurIPS*, 2019.

[4] H. Yin, A. Mallya, A. Vahdat, J. M. Álvarez, J. Kautz, and P. Molchanov, "See through gradients: Image batch recovery via gradinversion," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16 332–16 341, 2021.

[5] B. Zhao, K. R. Mopuri, and H. Bilen, "idlg: Improved deep leakage from gradients," *CoRR*, vol. abs/2001.02610, 2020. [Online]. Available: http://arxiv.org/abs/2001.02610

[6] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[7] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.

[8] J. Ball, "NSA collects millions of text messages daily in 'untargeted' global sweep," 2014. [Online]. Available: https://www.theguardian.com/world/2014/jan/16/nsa-collects-millions-text-messages-daily-untargeted-global-sweep

[9] D. J. Leith and S. Farrell, "Contact Tracing App Privacy: What Data Is Shared By Europe's GAEN Contact Tracing Apps," in *Proc IEEE INFOCOM*, 2021.

[10] D. J. Leith, "Mobile Handset Privacy: Measuring The Data iOS and Android Send to Apple And Google," in *Proc Securecomm*, 2021.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[12] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, p. 2222–2232, Oct 2017. [Online]. Available: http://dx.doi.org/10.1109/TNNLS.2016.2582924

[13] K. Zhang, X. Song, C. Zhang, and S. Yu, "Challenges and future directions of secure federated learning: a survey," *Frontiers of Computer Science*, vol. 16, no. 5, p. 165817, Oct. 2022. [Online]. Available: https://link.springer.com/10.1007/s11704-021-0598-z

[14] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," *CoRR*, vol. abs/2003.02133, 2020. [Online]. Available: https://arxiv.org/abs/2003.02133

[15] P. Kairouz, H. B. McMahan, and . Brendan Avent, "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, pp. 1–210, 2021.

[16] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2017.

[17] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 691–706, 2019.

[18] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," *2019 IEEE Symposium on Security and Privacy (SP)*, May 2019. [Online]. Available: http://dx.doi.org/10.1109/SP.2019.00065

[19] O. Press and L. Wolf, "Using the output embedding to improve language models," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 157–163. [Online]. Available: https://aclanthology.org/E17-2025

[20] S. L. Blodgett, J. T.-Z. Wei, and B. O'Connor, "Recognizing global social media english with u.s. demographic modeling," in *Proc Workshop on Noisy User-Generated Text*. Association for Computational Linguistics, 2017.

[21] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the study of sms spam filtering: new collection and results," in *DocEng '11*, 2011.

[22] A. Marzal and E. Vidal, "Computation of normalized edit distance and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 926–932, 1993.

Data Augmentation for Opcode Sequence Based Malware Detection

Niall McLaughlin and Jesus Martinez del Rincon

Centre for Secure Information Technologies (CSIT), Queen's University Belfast
{*n.mclaughlin, j.martinez-del-rincon*}*@qub.ac.uk*

Abstract

In this paper we study data augmentation for opcode sequence based Android malware detection. Data augmentation has been successfully used in many areas of deep-learning to significantly improve model performance. Typically, data augmentation simulates realistic variations in data to increase the apparent diversity of the training-set. However, for opcode-based malware analysis it is not immediately clear how to apply data augmentation. Hence we first study the use of fixed transformations, then progress to adaptive methods. We propose a novel data augmentation method – Self-Embedding Language Model Augmentation – that uses a malware detection network's own opcode embedding layer to measure opcode similarity for adaptive augmentation. To the best of our knowledge this is the first paper to carry out a systematic study of different augmentation methods for opcode sequence based Android malware classification.

1. Introduction

Data augmentation is used to improve the generalisation of machine-learning models by artificially increasing the diversity of the training data [1]. Data augmentation is needed because we often have limited training data, which does not encompass the full diversity of in-the-wild data. Using augmentation we can expose the network to a larger variety of data than is present in our limited training set. The use of augmentation has gained increasing importance due to deep-learning which often requires large amounts of training data. The disadvantage of data augmentation is that the augmented samples may be highly correlated with the existing training data, so this approach is necessarily limited in the performance boost it can provide. When applying data augmentation, the variability of samples in the original training set is artificially increased by modifying each training example using one or more transformation operations. The transformation operations are usually designed to mimic natural variations in the data. For instance, in image classification we may want an object detector to recognise a cat whether it is seen from the left or right, or rotated slightly. Hence, mirroring and/or small rotations are commonly used for vision tasks. The goal is to make the model invariant to these transformations, thus improving its robustness to similar real world variations. However not all augmentations correspond to real-world variations e.g. Mixup [2]. Nevertheless, such methods have been shown empirically to improve generalisation. For familiar types of data such as images, video, and audio, intuition can often provide guidance on the design of novel augmentation schemes. However, for more abstract data, such as opcode sequences, designing augmentation methods is more challenging. In this paper, we study data augmentation applied to opcode-sequence based malware classifiers [3], [4]. However, our methods are general, meaning they could be applied to sequences of raw bytes e.g., Malconv [5], or 2D malware images [6]. The contributions of this work are:

(i). We perform a systematic study of data augmentation methods applied to opcode-sequence based malware detection.

(ii). We introduce a novel adaptive data augmentation technique, Self-Embedding Language Model Augmentation, shown in Fig. 1. This method uses the network's own opcode embedding layer to measure opcode similarity to apply adaptive data augmentation during training.

2. Related Work

Opcode based malware detection has been extensively studied. Some of the early approaches were based on short n-grams [7]–[9] using classical machine learning algorithms. However they required extensive pre-processing and feature selection. Recently, deep-learning architectures designed for image classification have been applied [6], [10], where the opcode sequences is transformed into an image for classification by existing image classification network architec-

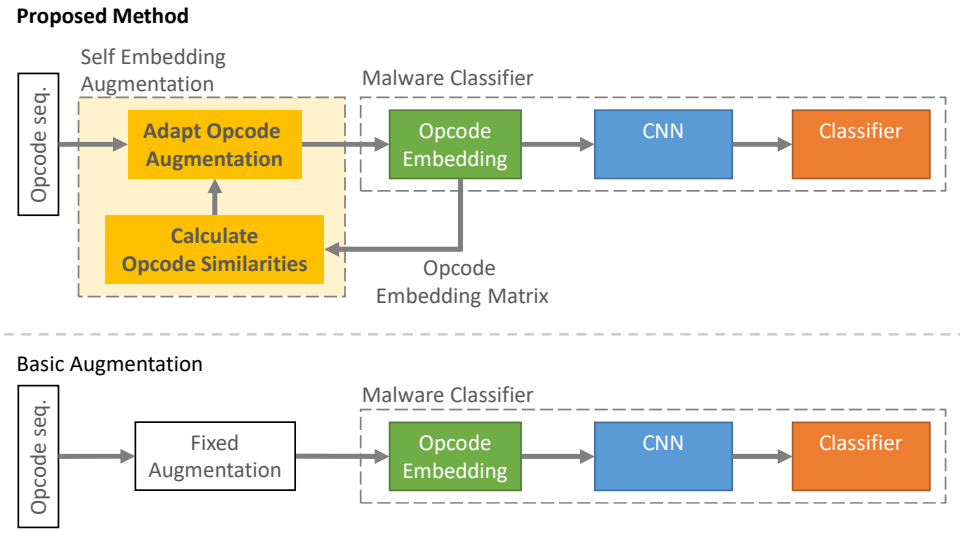


Figure 1: Top: Proposed Self-Embedding Augmentation for opcode-sequence based malware classifiers. The network's own opcode embedding matrix is used to generate realistically augmented versions of opcode sequences. Augmentation parameters are updated online during network training to generate better augmentations. Bottom: A basic augmentation approach that uses fixed parameters.

tures. Following from this, efficient sequence processing architectures originally proposed for natural language processing (NLP) [11] e.g., 1D CNNs, were used. By processing sequences using local filters, the inefficiency of recursive models such as LSTMs [12] is avoided. To the best of our knowledge, data augmentation has not been applied with such methods to date. This is likely due to the difficulty of designing an appropriate augmentation scheme for computer code.

Opcode-sequence based malware detection shares commonalities with NLP. Data augmentation methods in NLP include: synonym replacement, random word insertion, random word swap and random word deletion. While these techniques may give some benefit [13] they tend to work best on smaller datasets. In the masking technique words are replaced with a special 'blank' word [14]. This can be seen as a form of regularisation that works by adding noise to the input. A related method, replaces words using synonyms from a thesaurus [11] to create semantically similar but different sentences. Building on the idea of substituting semantically similar words, a word embedding model such as GloVe [15] can be used to measure word similarity. This allows replacement of words with semantically close neighbours based on their word embeddings [16]. The NLP techniques above are not easy to apply to opcode sequences as naive substitution of opcodes may break program functionality. There are also no widely available pre-trained embeddings for opcode sequences. In the field of malware classification we are aware of only one other related work, which applies additive noise to 3-channel encoded binary file images [17] for augmentation.

3. Method

3.1. Opcode Sequence Malware Classification

The opcode sequence of a given program can be recovered either via disassembly or by dynamic analysis. In this work we focus on static analysis by disassembly, although our methods are general enough to be applied to dynamic analysis. We use a Convolutional Neural Network (CNN) based malware classifier [3]. The complete network consists of an opcode embedding layer, a single convolutional layer, a max pooling layer, a linear classification layer, outputting the probability a sample is malicious. This style of architecture is often referred to as *Malconv* and has been successfully validated [3], [4], [18].

3.2. Data Augmentation of Opcode Sequences

We use the term data augmentation to describe the technique where the raw inputs to a network are modified during training. The inputs are usually modified to reflect variations the network should be invariant to during testing [1]. In a sense, data augmentation artificially increases the size of the training-set and is commonly applied to improve the network's generalisation performance by showing it a greater variety of examples than those present in the training-set.

Data augmentation can be applied either offline or online [1], [19]. In online augmentation every training example is randomly modified in a different way at every training epoch, thus increasing the variety of training data. Offline augmentation is typically applied

when the augmentation is too computationally costly to be used during the network's training loop [19]. For text/sequence based inputs, such as opcodes sequences, the computational costs of online augmentation are negligible, therefore in this work we always use online augmentation.

Typically, the strength of augmentation must be carefully selected so that the network converges and generalises well on real data. There are several ways in which the 'strength' of data augmentation can be varied. Firstly, the ratio of augmented data to un-augmented data presented to the network during training can be varied. We use the parameter β to describe the probability the network will be presented with an augmented sample at any given epoch. Varying β varies the weighting of the learned features between augmented and unagumented data.

Secondly, we can directly vary the strength of the augmentation methods themselves. This parameter will depend on the specifics of each augmentation technique. For example, we can vary the percentage of opcodes randomly replaced with zeros (See Section 3.3). We use parameter α to denote the strength of the augmentation.

Finally, we can vary whether to use data augmentation during training only, or whether to use test-time augmentation [20]. In test-time augmentation, multiple augmented samples are passed through the network at inference time and the final prediction based on aggregating the multiple predictions. Test-time augmentation can improve performance in some cases [20], [21]. In this work we only use data augmentation during training, which simplifies our analysis. Our reported performance figures therefore represent a lower bound and may be further improved by the use of test-time augmentation.

3.3. Basic Opcode Sequence Augmentation Methods

We will now introduce several basic methods of opcode sequence based data augmentation. These methods do not adapt to the context of a particular sequence, and are based on either heuristic principles, or existing approaches that have successfully been applied to text-based data augmentation in NLP [13], [14], [16].

Input Dropout. A random fraction of opcodes in each program are replaced with zeros. Opcodes for replacement are selected uniformly at random from the whole opcode sequence. This can be seen as a form of dropout [22] applied directly to the input. It is similar to methods that add noise to the input sequence to improve generalisation [14]. The strength of augmentation can be varied by varying the hyperparameter, α, which specifies the fraction of opcodes to randomly replace.

Note that in the case of opcode sequences, this operation is not the same as replacing opcodes with the `nop` (no operation) instruction. We reserve as a special 'blank' character used only for augmentation with zeros. This 'blank' character does not appear anywhere in the original opcode sequences, so it has no semantic significance in terms of the programmatic code, and instead simply represents missing information.

Random Replacement. A fraction of opcodes selected uniformly at random is replaced with different opcodes selected uniformly at random from the instruction set. The hyperparameter α specifies the fraction of opcodes to randomly replace, hence it represents the augmentation strength. This augmentation method increase the noise present in the opcode sequence and may force the network to learn to ignore outlier opcodes in the opcode sequence or to ignore short unrealistic opcode sequences.

Similar Instructions. A fraction of opcodes in each program are randomly replaced with an instruction selected from a hand-designed list of semantically similar instructions. Opcodes to be replaced are selected uniformly at random from the opcode sequence.

For every instruction in the instruction-set, a list of similar instructions was created using the names of instructions in the Android virtual machine instruction-set. For instance, instructions such as `move, move/from-16, move-16` etc. can be regarded as semantically similar. For each instruction, a table of similar instructions was created based on the instruction prefixes e.g.: `move, const, goto, cmp, if, get, put, cast` etc. Instructions without obvious semantic neighbours are augmented using the random replacement method. This augmentation technique is intended to create novel, but semantically similar, malware samples to ensure better generalisation of the classifier. The hyperparameter α specifies the fraction of opcodes in a given program to replace, hence it varies the augmentation strength.

Correlated Input Dropout. One possible issue with the above data augmentation methods such as input dropout, random replacement and similar instructions, is they produce uncorrelated variations in the input sequence. This is because opcodes to be replaced are selected uniformly at random. This method of augmentation may not create enough variation to pose a challenge to the classifier, and therefore may not significantly improve performance. To increase the strength and difficulty of the augmentation we instead examine the use of correlated augmentations.

Rather than replacing individual opcodes uniformly at random we instead examine correlated opcode replacement. We select one or more instructions from the instruction-set, then replacing all instances of these instructions in the opcode sequence with the special reserved 'blank' character. For example, if we select the move instruction, all instances of the move instruction in the opcode-sequence would be replaced with 'blank'. The same process can be repeated for additional instructions to increase augmentation strength.

This form of augmentation has a significant effect on the program semantics and may force the network to learn more robust features, as it cannot rely on any particular instruction in isolation to perform classification. It must instead learn more robust opcode patterns. This method differs from Input Dropout as it forces the network to learn with significant amounts of correlated noise. The parameter α controls the number of the instructions to replace as a faction of the total size of the processor's instruction-set, hence varies augmentation strength.

3.4. Adaptive Opcode Sequence Augmentation Methods

In this section we introduce several adaptive augmentation methods. These methods use knowledge of the semantic similarity of opcodes to augment samples in a way that may be more realistic than the basic basic augmentation methods introduced in Section 3.3.

3.4.1. Language model. First introduced for natural language processing, language models learn the semantic relationships between words in natural language text. A language model can be used to measure the semantic similarity of different words. Language models can be trained to predict a missing character/word in a sequence of text conditioned on the neighbouring characters/words. Alternatively,they can be trained to predict the next character/word in a sentence based on the preceding words [23]. We adapt this idea for opcode sequence based malware analysis. Given a large number of opcode sequences, we train an off-the-shelf language model to predict a missing opcode given its neighbouring opcodes. The resulting model captures the semantic similarity of different opcodes. We use the Continuous Bag of Words (CBOW) word2vec algorithm [24] as our language model. Given a word2vec model trained on many opcode sequences, its embedding matrix contains information on the semantic similarity of different opcodes [16]. Opcodes with similar embedding vectors tend to be semantically similar. We can then perform augmentation by replacing random opcodes in a given sequence with their semantic equivalents.

Concretely, word2vec model is trained offline using all the opcode sequences in the training-set. After training, we extract the word2vec opcode embedding matrix. For every opcode we create a ranked list of its most similar opcodes, based on the Euclidean distance between embedding vectors. For every opcode we record its top-10 most similar opcodes. To perform augmentation of a given input sequence, opcodes are selected uniformly at random from the sequence. Each opcode is then replaced with an opcode randomly selected from that opcode's top-10 list of semantically similar neighbours. The hyperparameter α specifies the fraction of opcodes in the original sequence to replace.

3.4.2. Self-Embedding Language Model. As part of training a *Malconv* like CNN to perform malware detection we are also learning an opcode embedding matrix, as described in Section 3.1. The opcode embedding matrix learns the semantic relationships between opcodes. We note that the opcode embeddings learned for malware classification may be different from those learned by a language model. Malware detection and language modelling are fundamentally different tasks and may require the network to extract different information from the opcode sequence, hence each opcodes may have different semantics in each task.

We therefore hypothesise that the embedding matrix from a network trained on malware classification may be more useful for generating realistically augmented inputs for training a malware classifier. This poses a chicken-and-egg problem where we require a trained malware classifier in order to train a new one. We resolve this problem by using the embedding matrix from the network currently being trained. We use this embedding matrix to measure the semantic similarity of opcodes and hence generate realistically augmented training samples. As the network trains, the embedding matrix used to generate augmented samples is updated, hence the augmented samples become more realistic and challenging as training progresses. To allow the network time to adapt to the increasingly challenging augmented samples we use a lagging version of the embedding matrix, updated once per training epoch.

At the start of every training epoch, a copy of the malware detection network's opcode embedding matrix is made. Given this opcode embedding matrix, for every opcode a list of semantically similar opcodes is constructed based on the Euclidean distance between embedding vectors. For every opcode, the top-10 most semantically similar opcodes i.e., those with smallest Euclidean distances, are recorded to produce an opcode similarity table. To perform augmentation of a given opcode sequence during training, a fraction of the sequence's opcodes are selected uniformly at random for replacement. The hyperparameter α specifies the fraction of opcodes to randomly replace. Each selected opcode is then replaced with a opcodes, selected uniformly at random, from the original opcodes lists of semantically similar neighbours. The complete augmentation process is illustrated in Fig. 1.

We note that when training first begins, the embedding matrix is randomly initialised, so there is no semantic information yet encoded in its weights. Hence this method essentially performs random opcode replacement during early training epochs, and gradually begins to incorporate semantic information as training progresses. This simulates a curriculum based approach and means that the strength of the augmentation gradually increases during training.

4. Experiments

To test the effectiveness of our augmentation algorithms we performed experiments using several different malware datasets and with different network hyper-parameters. For all experiments we use the 1D CNN network design from [3], which is similar to that of [5]. The following hyper-parameters were used: Embedding Layer: 8 dimensions, Convolutional filters: 64, of length 8, Max-Pooling layer, followed by a single linear layer with a 1 dimensional output. Binary cross entropy loss was used. Adam optimiser [25] with batch size 48, and learning rate of 1e-3 for 120 epochs. During training and testing opcode sequences were truncated to $128,000$ opcodes due to GPU memory limitations. For all experiments, all code and hyper-parameters remained constant across datasets and models, with only the data augmentation method varying. Results are reported using f1-score.

Two datasets were used for the experiments: The Small Dataset, which consists of malware from the Android Malware Genome project. This dataset contains 2123 applications - 863 benign and 1260 malware samples from 49 malware families. The Large Dataset was provided by McAfee Labs (Intel Security) and consists of malware from the vendor's internal dataset. This dataset contains roughly 10,000 malware and a further 10,000 clean applications collected from the Google Play Store. Both datasets are therefore quite evenly balanced in terms of malware and clean programs.

Both datasets were thoroughly checked and cleaned to ensure no duplicate programs that contaminated the test/training splits. For experimentation purposes, each dataset was split into 5 non-overlapping folds and cross validation was preformed during evaluation. In effect, during each round of cross validation, 80% of the dataset was used for training and 20% for testing. For each experiment we report the average performance across the 5 cross-validation folds in terms of f1-score.

In our experiments we vary the α parameter, which controls the augmentation strength, and report how this affects the classification f1-score. For all experiments we hold the parameter β, which controls the probability of applying augmentation to a given sample, constant at $\beta = 0.5$ i.e., half the samples presented during training are augmented and half are unchanged. While it would be be possible to vary both parameters for all experiments, we did not explore this possibility due to the excessive computational costs.

For the language model augmentation method we use the Gensim implementation [26] of word2vec [24]. We use the default parameters i.e., a continuous bag-of-words (CBOW) model trained using negative sampling. We use word2vec embedding vectors of dimensionality 8 in order to match the embedding vector size of the malware detector network. The context window size is 5 and the model was trained for 5 epochs. For each experiment the word2vec model was pre-trained offline using the same training data as the main malware classifier. The embedding matrix was then extracted and used to build an opcode similarity table for generating augmentations.

4.1. Baseline Performance

Baseline performance of the 1D CNN opcode-based malware detector network, with no data augmentation, was measured on both the Small and Large datasets. The average f1-score from 5-fold CV was recorded for each dataset. For all remaining experiments, all network hyper-parameters, training and testing settings and code, except that related to data augmentation, remained identical. Baseline results are shown in Table 1.

4.2. Basic Augmentation Methods

We compare the basic augmentation schemes with baseline no augmentation results. The basic augmentation methods make predefined changes to the opcode sequence without regard to context. They include: 'Input Dropout', 'Random Replacement', 'Similar Instructions' and 'Correlated Input Dropout' (See Section 3.3). The augmentation strength, α, was systematically varied. Note that the α parameter controls different aspects of the strength of each augmentation method, so we cannot directly compare α values across different augmentation methods.

The results of these experiments are shown in Table 1. Across both datasets, several augmentation methods produce consistent improvements over the baseline for a range of α values. In particular, Input Dropout consistently performs well with peak performance occurring at $\alpha = 0.2$ on both datasets. Several other methods, Random Replacement and Similar Instructions, outperform Input Dropout on the Small Dataset. However this performance boost is not repeated on the Large Dataset. Correlated Input Dropout performs similarly across both datasets, although its overall performance is slightly worse than Input Dropout in terms of relative improvement compared to the baseline across both datasets. The results overall suggest that these basic augmentation methods have the potential help to improve malware classification performance.

4.3. Language Model Augmentation

This approach generates augmented opcode sequences while aiming to preserve functionality and semantics. The gensim [26] implementation of CBOW word2vec [24] with 8-dimensional opcode embedding vectors was used to calculate opcode similarity. The language model was trained offline and its embedding matrix extracted to calculate opcode similarities.

The results of the experiments are shown in Table 1. Performance on the Small Dataset is comparable with

the Self-Embedding (SE) Language Model (see Section 4.4) method however performance on the larger dataset is similar to the basic augmentation methods. We hypothesise that the relatively poor performance on the Large Dataset, compared to the SE Language model, may be caused by the fact that this language model is trained on a different task from malware analysis i.e., the language model is trained to predict an opcode from its context. The relationships between opcodes learned by the language model may be different those that are important for malware classification. However the overall performance of this method is better than or competitive with any of the other augmentation methods. Improving its performance may therefore present an avenue for future research.

4.4. Self-Embedding Language Model Augmentation

In Self-Embedding language model augmentation, the semantic similarity between opcodes for data augmentation is measured using the constantly updated embedding matrix from the malware classifier currently being trained. In these experiments the network's embedding matrix was sampled at the beginning of every epoch. Hence the table of opcode similarity is updated one per epoch.

We compare performance against the baseline system with no augmentation, and against all the other augmentation methods. Results are reported in Table 1. For both the Large and Small datasets we can see that augmentation by self-embedding language model has a positive effect on classification performance. Across both datasets the effectiveness of augmentation varies as the α value is varied. In both cases maximum performance occurs when α is around 0.2. Performance drops off at both higher and lower α values. Compared to the other augmentation methods we can see that this method has the highest performance across both datasets. When the optimal α value is selected, the f1-score increases by almost 1% which is the largest improvement seen across all the augmentation methods. We can also see that this method consistently outperforms the baseline, showing that the value of the α parameter is not critical. We propose this method should be used whenever training opcode-sequence based malware classifiers that make use of an embedding matrix.

4.5. Augmentation Performance in Context

In this section we provide context for the performance gains achievable using data augmentation. Another way to improve the network's performance is to increase its size until just before over-fitting occurs. This comes at the cost of increased training and inference time, however it provides an upper-bound for the potential performance of the model architecture on the dataset. We therefore compare the performance of a constant sized network trained using data augmentation, with that of increasingly larger networks trained without data augmentation.

On the Large Dataset We train networks with between 32 and 256 convolutional filters without using data augmentation. We compare these networks against a 64 convolutional filter network trained using all proposed methods of data augmentation with their optimal hyperparameters. The results are shown in Fig. 2. We do two types of comparison: firstly, comparing classification performance as model size is varied, and secondly, comparing classification performance versus training and inference time.

Fig. 2 (a) shows that for models trained without data augmentation it is necessary to significantly increase the parameter count to increase performance. In contrast, by using data augmentation a networks with 64 convolutional filters can approach or exceeds the performance of a network with 128 filters but no augmentation. The self-embedding language model augmentation method even allows the smaller network to slightly exceed the performance of the 128 filters network trained without augmentation.

Fig. 2 (b) shows that increasing network size significantly increases both training and inference time. However for a constant network size data augmentation causes little change in training or inference time. Augmentation enables the performance the 64 convolutional filter network to slightly exceed the performance of the larger 128 filter network trained without augmentation. In addition, the smaller network achieves this with much lower training and inference time.

These experiments show that our proposed augmentation methods can significantly improve malware classification performance, without altering the memory or computation time needed. They allow a smaller network trained using augmentation to exceed the performance of a network with no augmentation.

4.6. Combinations of Augmentations

In this section we study the performance of combinations of augmentation methods. Using different augmentation methods together introduces more variability into the training data, which may improve generalisation. Note that we did not modify the augmentation methods to prevent interference or provide knowledge of previous augmentations applied. As above, a network with 64 convolutional filters was trained using various combined augmentations. In Fig. 3 we report results over a range of augmentation strengths. Note that due to the computational costs of these experiments, only a small number of combinations were tried. We selected combinations of augmentations based on those that performed well individually. These experiments were performed on the Large Dataset.

Method	Input Dropout		Random		Similar		Correlated		Lang. Model		Self-Embed	
Dataset	Small	Large	Small	Large	Small	Large	Small	Large	Small	Large	Small	Large
Baseline	0.94939	0.94363	0.94939	0.94363	0.94939	0.94363	0.94939	0.94363	0.94939	0.94363	0.94939	0.94363
α 0.05	0.94877	**0.94537**	**0.95688**	**0.94780**	**0.95549**	**0.94721**	**0.95188**	**0.94794**	**0.95835**	**0.94866**	**0.9502**	**0.94666**
α 0.1	**0.95176**	**0.94689**	**0.94979**	**0.94595**	**0.95145**	**0.94721**	**0.9529**	**0.94799**	**0.94990**	**0.94829**	**0.95118**	**0.94926**
α 0.2	**0.95539**	**0.94918**	**0.95374**	**0.94647**	**0.95063**	**0.94604**	**0.95406**	**0.94860**	0.94129	**0.94731**	**0.95881**	**0.95061**
α 0.3	**0.94994**	**0.94739**	0.94049	0.94286	0.94306	0.94350	**0.95220**	**0.94650**	0.94346	**0.94434**	**0.9576**	**0.94873**
α 0.4	0.94697	**0.94681**	0.94155	0.93952	0.94784	0.94353	**0.95100**	0.94353	0.93547	0.94085	**0.95032**	**0.94775**
α 0.5	0.94487	0.94274	0.93640	0.93551	0.94018	0.94045	0.94854	0.94241	0.93510	0.93908	**0.95267**	**0.94827**
Max	0.95539	0.94918	0.95688	0.94780	0.95549	0.94721	0.95406	0.94860	0.95835	0.94866	0.95881	0.95061
Delta	0.00600	0.00555	0.00749	0.00417	0.00610	0.00358	0.00467	0.00497	0.00896	0.00503	0.00942	0.00698

TABLE 1: Comparison of all augmentation methods as the augmentation strength (α) is varied between 0.05 and 0.5. Results are reported as average f1-score from 5-fold cross-validation across Large and Small Datasets. Entries in **bold text** denote improved performance compared to the baseline. Max - The maximum absolute f1-score. Delta - The maximum absolute improvement over all α values, compared to the baseline.

[5] E. Raff, J. Barker, J. Sylvester, R. Brandon, B. Catanzaro, and C. K. Nicholas, "Malware detection by eating a whole exe," in *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[6] R. Kumar, Z. Xiaosong, R. U. Khan, I. Ahad, and J. Kumar, "Malicious code detection based on image processing using deep learning," in *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*, 2018, pp. 81–85.

[7] G. Canfora, F. Mercaldo, and C. A. Visaggio, "Mobile malware detection using op-code frequency histograms," in *2015 12th International Joint Conference on e-Business and Telecommunications (ICETE)*, vol. 4. IEEE, 2015, pp. 27–38.

[8] B. Kang, B. Kang, J. Kim, and E. G. Im, "Android malware classification method: Dalvik bytecode frequency analysis," in *Proceedings of the 2013 research in adaptive and convergent systems*, 2013, pp. 349–350.

[9] Q. Jerome, K. Allix, R. State, and T. Engel, "Using opcode-sequences to detect malicious android applications," in *ICC*. IEEE, 2014, pp. 914–919.

[10] R. U. Khan, X. Zhang, and R. Kumar, "Analysis of resnet and googlenet models for malware detection," *Journal of Computer Virology and Hacking Techniques*, vol. 15, no. 1, pp. 29–37, 2019.

[11] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in Neural Information Processing Systems*, 2015, pp. 649–657.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *(EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019.

[14] Z. Xie, S. I. Wang, J. Li, D. Lévy, A. Nie, D. Jurafsky, and A. Y. Ng, "Data noising as smoothing in neural network language models," in *ICLR*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=H1VyHY9gg

[15] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[16] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 4163–4174.

[17] F. O. Catak, J. Ahmed, K. Sahinbas, and Z. H. Khand, "Data augmentation based malware detection using convolutional neural networks," *PeerJ Computer Science*, vol. 7, p. e346, 2021.

[18] S. Millar, N. McLaughlin, J. M. del Rincon, and P. Miller, "Multi-view deep learning for zero-day android malware detection," *Journal of Information Security and Applications*, vol. 58, p. 102718, 2021.

[19] T. Chilimbi, Y. Suzue, J. Apacible, and K. Kalyanaraman, "Project adam: Building an efficient and scalable deep learning training system," in *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, 2014, pp. 571–582.

[20] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag, "When and why test-time augmentation works," *arXiv preprint arXiv:2011.11156*, 2020.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[23] Y. Bengio, "Neural net language models," *Scholarpedia*, vol. 3, no. 1, p. 3881, 2008, revision #140963.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, 2010, pp. 45–50.

Convolutional Neural Network for Software Vulnerability Detection

Kaixi Yang
Centre for Secure Information Technologies(CSIT)
Queen's University Belfast, eBay
Belfast, United Kingdom
kyang03@qub.ac.uk

Paul Miller
Centre for Secure Information Technologies (CSIT)
Queen's University Belfast
Belfast, United Kingdom
p.miller@qub.ac.uk

Jesus Martinez-del-Rincon
Centre for Secure Information Technologies (CSIT)
Queen's University Belfast
Belfast, United Kingdom
j.martinez-del-rincon@qub.ac.uk

Abstract—**Exploitable vulnerabilities in software are one of the root causes of cybercrime, leading to financial losses, reputational damage, and wider security breaches for both enterprise and consumers. Furthermore, checking for vulnerabilities in software is no longer a human-scale problem due to code volume and complexity. To help address this problem, our work presents a deep learning (DL) model able to identify risk signals in Java source code and output a classification for a program as either vulnerable or safe. Sequences of raw Java opcodes are used to train a convolutional neural network that automatically encapsulates discriminative characteristics of a program that are then used for the prediction. Compared to traditional machine learning methods, this approach requires no prior knowledge of the software vulnerability domain, nor any hand-crafted input features. When evaluated on the publicly available benchmark dataset Juliet Test Suite containing 38520 vulnerable and 38806 safe programs, our method achieves an F1 score of 0.92.**

Keywords—***Software Vulnerability, Deep Learning***

I. INTRODUCTION

Detecting hidden flaws in software is an important and challenging problem. Failure of detection during production and review could result in attacks which take control of the system. Hence, the software industry is paying increasing attention to the robustness of software. Many vulnerabilities are reported in the Common Weakness Enumeration (CWE) list to help us record and classify these vulnerabilities. In 2021 alone, 918 CWEs were recorded. Vulnerability detection techniques include manual discovery, computer assisted discovery and fully automated discovery, with a growing shift towards latter two in recent years. However, existing tools can only detect limited types of vulnerabilities based on pre-defined rules [1].

This paper aims to investigate the use of DL techniques for the automatic detection of software vulnerabilities. We will focus on 112 different types of CWEs. The novelty of this project lies in developing a DL model that can automatically detect and learn weakness patterns in the source code, called risk signals, and use them as well as their relationships, to detect vulnerable code.

Our paper makes three contributions: First, we design and apply a model to analyse and detect Java source code vulnerabilities from the Juliet Test Suite [2]. Second, we investigate how to utilize a Convolutional Neural Network (CNN) to predict software vulnerabilities. Finally, we work with Java opcode sequences directly.

II. RELATED WORK

With the surge in deep DL, neural networks architectures have been applied to automatic code analysis [3]. This is especially notable on malware analysis [4], where CNNs [5], recurrent neural networks (RNN) architectures [6], as well as NLP techniques[7], have all obtained promising results. However, little research exists on the application of DL for software vulnerability detection [1][8]. Russell et al.[1] leveraged a DL approach, including CNN and RNN, on open-source code from the Juliet Test Suite [2], Debian [9] and GitHub [10] repositories to combine synthetic code snippets with natural code. They used word2vec to get a source code embedding representation then trained this representation on CNN and RNN. The result showed that CNN trained on the Juliet Test Suite data performed much better than natural functions from Debian and GitHub because of the coding structures and styles. Different from working with source code [1], authors in [8] proposed a model called Instruction2vec which vectorizes the instructions of assembly code efficiently. The vectors include opcodes and operands. Then they fed the vectorized assembly code to a text-convolutional neural network for training. The model achieves up to 96.1% accuracy in classifying whether the function has software weaknesses.

III. METHOD

A. Pre-processing of Source Code

The Juliet Test Suite data is composed of Java files containing examples of both robust and vulnerable code. As noted, in the introduction, rather than using the raw source code, our system will input the bytecode sequence. Thus, each Java file is first compiled into bytecode using javap [11]. Each bytecode is composed of one byte that represents the opcode, along with zero or more bytes for operands. In our case, we only use opcodes for experiments and discard the operands. We extract the opcode sequence from each class file by using Java Bytecode Instructions list [12]. The preprocessing will provide one opcode sequence for each of the classes. The preprocessing of Juliet Test Suite data is shown in Fig. 1, we can link back opcode *bb* to *new* on line 3 in the bytecode.

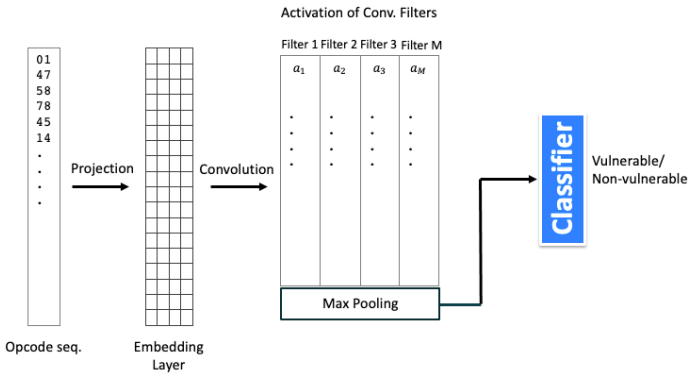


Fig. 2: Deep learning architecture for software vulnerability detection.

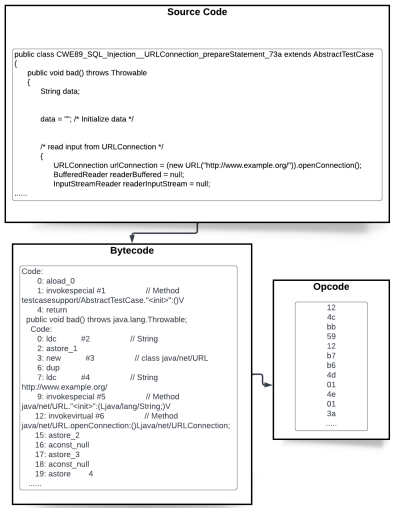


Fig. 1: Workflow of how a Java file is disassembled to produce an opcode sequence

B. *Network Architecture*

Fig. 2 depicts our proposed DL architecture. This model uses the opcode sequence as input to predict the label for the given class file as either vulnerable or not. Note that our network only uses one convolutional layer.

1) Opcode Embedding Layer

Opcodes which are semantically similar will have similar representations in the embedding. This concept is inspired by [4]. In our system, we encoded a sequence of opcode instructions as one-hot vectors. To form a one-hot vector we matched each of opcode with an integer between 1 and D, in our case we defined 204 opcodes, $D = 204$. Then the opcode sequences are project into an k-dimensional embedding space by multiplying each sequence by weight matrix $W \in R^{D \times k}$, where k is the dimensionality of the embedding space. The result of this projection can be represented by a matrix P of size $n \times k$, where n is the n'th opcode in the sequence and each row in P corresponds to the representation of each opcode sequence.

2) Convolutional Layers

The convolutional layer learns to detect the risk patterns in the opcode sequences. It is represented by a matrix of weights with which we convolve the opcode sequences. At each position in the opcode sequence, we take the product of the convolutional filter weights with the embedding layer values and sum the result. We can apply many convolutional filters to the opcode sequence and record the response (or activation) of each filter at each location in the opcode sequence. The weights in the filter are trainable, using backpropagation. During the learning process, the network will find the best filters.

The convolution layer takes the embedding matrix P as input. Note that our convolutional layer has one filter with size of $s \times k$, the width of filter should be the same as the dimension of our embedding layer. Passing samples through a convolutional layer, each of the filters produces an activation map a, which can be stacked to produce a matrix A, of size $n \times m$. The convolution of the filters with program embedding matrix P can be denoted as:

$$a = ReLU(Conv(P)w_m b_m) \quad (1)$$

and the output matrix from activation function A is donated as:

$$A = [a_1|a_2| \dots |a_m] \quad (2)$$

where w_m and b_m are the respective weight and bias parameters of the m'th convolutional filter of convolution layer, where *Conv()* represents the mathematical operation of convolution operation of the filter with the input, and where the activation function $ReLU(x) = \max\{0, x\}$is used.

Given the output matrix from the convolutional layer, max pooling is then used over the program length dimension to give a vector f containing the maximum activation of each convolutional filter over the program length. The main idea behind max pooling in this context is to allow us to deal with inputs of any length. The output contains the most prominent activation of the previous convolution output .

Then we perform dropout on the output matrix from the max-pooling layer. We do this because dropout can prevent our model from overfitting in the training phase. During dropout, some neurons in our network were omitted with a random probability r. where r is a vector of independent Bernoulli random variables each of which has probability of being 1. With dropout, the output is:

$$f = [\max(a_1 * r)|\max(a_2 * r)| \dots |\max(a_m * r)] \quad (3)$$

3) Classification Layers

Finally, the feature vectors from the dropout layer are concatenated into a single feature vector f, which is passed to a multilayer perceptron (MLP) consisting of a fully connected hidden layer and a fully connected output layer. Adding a fully connected layer is a (usually) cheap way of learning non-linear combinations of the high-level features as

represented by the output of the convolutional layer. We can write the hidden layer as follows:

$$z = ReLU(W_h f + b_h) \quad (4)$$

where W_h, b_h, are the parameters of the fully connected hidden layer and activation function *ReLU* has been used again. Finally, the output z from MLP is passed to the Softmax classification layer, a Softmax layer normalized the output from fully connected layer to a probability distribution to classify which class the current sample belongs to, which gives the probability that code is vulnerable, denoted as follows:

$$p(y = i|z) = \frac{e^{z_i}}{\sum_{i'=1}^{I} e^{z_{i'}}} \quad \text{for } i = 1...I \quad (5)$$

where i is the class, and y is the label that indicates whether the sample is vulnerable or not. Our case is a two-class problem (vulnerable/ non-vulnerable), i.e., $I = 2$ and z is a two-element vector.

C. Learning Process

Our network's learning process is a method which improves the network's performance by updating its parameters. The job of the algorithm is to find a set of internal model parameters that performs well against some performance measure such as algorithm loss or error. In our case, the loss function can be described as the following:

$$Loss = -\sum_i^I Y \log\big(p(y = i|z)\big) \quad (6)$$

Y is the ground truth label. We aim to minimize the loss function and which is achieved with optimizers. Optimizers allow us to iteratively change the network weights and the learning rate of neural network to reduce the losses. This translates on investigating how the error changes as each weight changes. We can update the weights by using:

$$W^{new} = W - \alpha \frac{\partial \text{ Loss}}{\partial \text{ W}} \quad (7)$$

being W the weights across all the layers, $\frac{\partial \text{ Loss}}{\partial \text{ W}}$ the gradient or the direction of steepest measure, α the learning rate. The bigger the learning rate, the greater the change to the weights. This learning process is performed by stochastic gradient descent, which means that the parameters are updated after every batch of sample using the gradient of the loss function with respect to these parameters (eq. 7). This rule is applied repeatedly over the network until the parameters converge.

IV. RESULTS

Our experiments were carried out using the publicly available Juliet Test Suite [2]. The Juliet Test Suite test cases were created for use in testing static analysis tools. It is made up of synthetic code snippets and contains folders, within which there are multiple source code files, each containing a collection of deliberate errors corresponding to a specific CWE. Code examples with security vulnerabilities are given in simple form as well as embedded in variations of different control flow and dataflow patterns. Each CWE entry describes a class of security errors. For example, CWE306 describes 'Missing Authentication for Critical Function". Each class has methods of the same functionality well, and poorly, implemented. We define a sample for our model as one of those methods with/without a vulnerability. In this project, the data we used is a collection of test cases in the Java language and it covers 112 different CWEs. We use 77326 samples in total including 38520 vulnerable code method samples and 38806 of non-vulnerable samples.

The dataset was split into 80% for training, 10% for validation, and the remaining 10% kept separate for testing. It was important to ensure the ratio of clean to vulnerable samples was the same in the validation and testing as it was in the entire dataset. Results are reported by using loss, accuracy, precision (P), recall (R) and F1-Score and are calculated as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

$$P = \frac{TP}{TP+FP} \quad (9)$$

$$R = \frac{TP}{TP+FN} \quad (10)$$

$$F1 - Score = 2 \cdot \frac{P \cdot * R}{P+R} \quad (11)$$

Accuracy is a ratio of correctly predicted observations to the total observations. Precision is the proportion of positive identifications that were correct. Recall is the proportion of actual positives that were identified correctly. In our case, positive identifications are equivalent to a vulnerable program and negative identifications represent the non-vulnerable program. Our focus is on the performance of F1-Score since it is considered an overall metric of the performance of the system. The F1-Score is defined as the harmonic mean of the model's precision and recall.

A. Effect of Hyperparameters

1) Investigation into the Effect of the Filter number

In this experiment, we investigate the number of filters and their effect on performance. The reason for this tuning is to optimize the computational expense as well as the metrics. We set filter size to 3 and used the following number of filters: 1, 4, 8, 16, 32, 64 and 128. From Fig. 3, it can be concluded that adding more filters does not significantly improve performance after at least four filters were used. This is most likely due to overfitting, since the number of parameters in the model increases significantly with each additional filter. Therefore, we chose a filter number of 4.

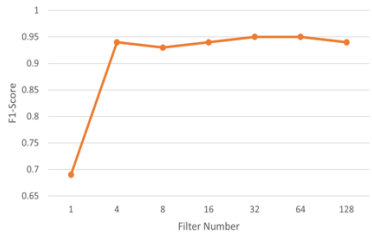


Fig. 3: Effect on Validation F1-Score with 1-128 filter numbers

2) Investigation into the Effect of the Convolutional Filter size

In this experiment, the filter size in the model was investigated. The size of the filters plays an important role in finding the key features. Thus, there is a need to determine the most suitable size of the filter. We experimented with filter sizes of 1, 3, 5, 7, 9, 11 and 13. From Fig. 4, we can see there is a significant performance improvement from a filter of 1 to 3. However, above 3 the performance only increases marginally with respect to filter size. Hence, a filter size of 3 was selected as a good balance between F1 performance and computational efficiency.

B. Vulnerability Detection Performance

We conducted a comparison between our model and two other state-of-art models [1, 8]. These papers developed DL models to detect vulnerabilities from C /C++ Juliet Test Suite data. We simply report their results in their papers obtained with different training and testing data to ours. So while this is not a like-for-like comparison, it provides reasonable insight.

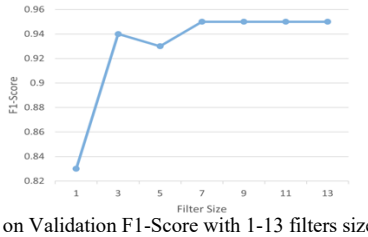


Fig. 4: Effect on Validation F1-Score with 1-13 filters sizes

The approach in [8] choses 3474 CWE-121 Stack-based Buffer Overflow functions as testcases and the examples are labelled as 'good case' and 'bad case'. TABLE I shows that [8] reached an accuracy of 0.96 by including opcodes and operands, such as register, point values and library function. Our system achieves very similar results without parsing the operands on a much larger dataset. Our method is effective across more than one CWE, whilst the evaluation by Jun Lee *et al* [8] is significantly limited due to only using one CWE.

The CNN architecture in [1] is similar to ours. At the data preparation stage, they removed more than 11,000 potential duplicated functions. After this removal, the data they used contains 11,896 examples from 118 CWEs, including 6,503 non-vulnerable cases and 5,393 vulnerable cases. TABLE I shows that they achieved an F1-score of 0.84. Our model performs better than [1], which is likely to be due to overfitting, since they removed duplicated functions and have less data than us.

TABLE I. Comparison of performance using our model vs. state-of-art model

	Training samples	**Testing samples**	**Number of CWE**	**F1-Score**	**Accuracy**
Our Model	61861	7733	1- 112	**0.92**	0.92
[1]	9517	1190	118	0.84	N/A
[8]	2779	695	1	N/A	**0.96**

V. CONCLUSIONS

In this paper we propose a CNN model for automatic vulnerability detection on Java code. We have conducted extensive experiments applying DL on a large public dataset and managed various studies about the effects of hyper parameters in order to inform future studies. We achieved performance comparable with the state of the art with a small and simple architecture, which minimizes the risk of overfitting. We found that for all convolutional filters added after the first filter, performance does not improve significantly with increasing number of filters. Varying the convolutional filter size also did not significantly improve model performance. Our model effectively identified vulnerabilities across more than one CWEs by using opcodes. However, our training, validation and testing datasets do not have the same proportion of each CWE, we cannot prove that our model can effectively detect vulnerabilities across 112 CWEs.

Future work should study the Java Juliet Test Suite further to access its quality. We intend to apply stratified k-fold [13] to ensure that training, validation and testing datasets contain approximately the same percentage of each CWE. Moreover, further analysis may show the performance per CWE and explore more of the 918 CWEs other than the 112 CWEs we have already worked with. As we only used 112 CWEs, we do not know how the model will perform on other CWEs. We will also focus on natural code datasets, for instance the Debian and GitHub datasets and data from eBay internal repos. According to a recent paper's [14] analysis, Grahn *et al* found the possible flaws from C/C++ Juliet Test Suite due to data leakage. Real-world code would allow the DL model learns vulnerabilities with more consistent and complex style and structure. Additionally, we will also consider using C/C++ test cases on the DL model for a better comparison with other models.

REFERENCE

[1] R. L. Russell *et al.*, "Automated Vulnerability Detection in Source Code Using Deep Representation Learning," *2018 17th IEEE Int. Conf. Mach. Learn. Appl. (pp. 757-762). IEEE.*, 2018.

[2] "Software Assurance Reference Dataset." https://samate.nist.gov/SARD/testsuite.php , 2022

[3] S. Wang *et al.* , "Automatically Learning Semantic Features for Defect Prediction," *2016 IEEE/ACM 38th Int. Conf. Softw. Eng. (pp. 297-308). IEEE.*, 2016.

[4] N. McLaughlin *et al.*, "Deep android malware detection," in *CODASPY 2017 - Proceedings of the 7th ACM Conference on Data and Application Security and Privacy*, Mar. pp. 301–308, 2017

[5] L. Mou *et al.*, "TBCNN: A Tree-Based Convolutional Neural Network for Programming Language Processing," *arXiv Prepr. arXiv1409.5718.*, 2014.

[6] Y. Xiao *et al.*, "Improving Bug Localization with Character-Level Convolutional Neural Network and Recurrent Neural Network," *Proc. - Asia-Pacific Softw. Eng. Conf. APSEC*, vol. 2018-December, pp. 703–704, 2018

[7] H. Ruan *et al.*, "DEEPLINK: Recovering issue-commit links based on deep learning," *J. Syst. Softw.*, vol. 158, 2019

[8] Y. Jun Lee *et al.*, "Learning Binary Code with Deep Learning to Detect Software Weakness.", 2017

[9] "Debian -- The Universal Operating System." https://www.debian.org/ , 2022

[10] "GitHub: Where the world builds software · GitHub." https://github.com/ , 2022

[11] "javap - The Java Class File Disassembler." https://docs.oracle.com/javase/7/docs/technotes/tools/windows/javap.html , 2022

[12] "Java byetecode instruction "http://en.wikipedia.org/wiki/Java_bytecode_instruction_listings, 2022

[13] "StratifiedKFold" https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html , 2022

[14] D. Grahn *et al.*, "An Analysis of C/C++ Datasets for Machine Learning-Assisted Software Vulnerability Detection," 2021.

Employee Cyber-Security Awareness Training (CSAT) Programs in Ireland's Financial Institutions

Mr. Reda Jouaibi
Faculty of Business & Hospitality
Technological University of The Shannon
Athlone, Ireland
A00277271@student.AIT.ie

Dr Aisling Keenan Gaylard
Faculty of Business & Hospitality
Technological University of The Shannon
Athlone, Ireland
aisling.keenan@tus.ie

Dr Brian Lee
Software Research Institute
Technlogical University of The Shannon
Athlone, Ireland
brian.lee@tus.ie

Abstract— ***This paper presents work in progress to analyse the use and effectiveness of cyber-security awareness training (CSAT) programs adopted in Ireland's financial institutions. The findings to date demonstrate that some Irish financial institutions provide CSAT programs for their employees, but their adoption and usage by employees and effectiveness remains unsure. We conclude that further research is required of CSAT effectiveness and whether these programs alleviate cyber threats.***

Keywords—Cyber-Security, Employee Awareness, CSAT Programs, Employee Training, Cyber Psychology

I. Introduction

The increase of technology advancement such as 5G along with the free internet access to hacking tools and forums have pushed Irish businesses and government institutions to invest more than ever before in cyber-security defence [1]. Over 50% of Irish businesses reported cyber-attacks in 2020 [2]. Financial organisations experience 300 times more cyber-attacks than organisations in different sectors with over 60% of large financial organisations suffering data breaches in 2020 [3]. Consequently, cyber-defences remain a challenge due to the ongoing expansion and the complexity of cyber-attacks [4]. And while humans are noted as the most vulnerable factor in an organisation's cyber-security systems [5], businesses should not only deliver cyber-security training to their workforce but, should also implement a more sustainable security awareness-based culture [6]. In the workplace context, the less experienced, trained, and least cyber-security aware users are the most vulnerable to cyber-attacks [7]. In response to this, IT experts along with government institutions such as the National Cyber-Security Centre (NCSC) advised and directed financial institutions to implement CSAT programs [8]. Today, CSAT programs have evolved and became part of most financial organisations' security and risk management processes [9].

From a theoretical perspective this research adopts two theories, the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT) to determine factors affecting technology usage towards positive cyber-security behaviour in the workplace. Moreover, this research will look at the impact of human factors such as motivation, personality traits, and social values on CSAT programs compliance in Ireland's financial institutions.

II. Research question & Objectives

The main objective of this research is to determine whether financial institutions in Ireland provides CSAT programs for employees and investigate their effectiveness. Therefore, the research questions are as follow:

- What makes an effective CSAT program?
- How to influence employees to comply and participate in CSAT programs?

III. Literature Review

A. CSAT Programs for Financial Institutions

CSAT programs are adopted by IT experts, government institutions, and organisations to help prevent and minimise cyber-risks. These programs are implemented to allow employees better understand their part in defending sensitive data from malicious attacks [10]. Moreover, CSAT programs familiarises employees with cyber-attacks they might face such as phishing emails, fraud, DDoS, or ransomware attacks [11]. Furthermore, the Central Bank of Ireland has advised the management of financial institutions to seriously consider the threat of cyber-threats and identify ways towards minimising such threats through risk management frameworks, and training and awareness programs [12].

B. Employees CSAT Programs for Financial Institutions

To mitigate cyber risks caused by humans, businesses along with financial institutions are adopting CSAT programs. Such programs improve employees' awareness as to why sensitive data must be secured from cybercriminals and ways to protect it [14]. Several scholars have determined the effectiveness and the benefits of CSAT programs for banks and organisations [15]– [17]. For instance, [17] suggest that banks and financial institutions must enhance their employees' awareness and cyber training programs for a better cyber-threat response. Furthermore, [18] indicated that CSAT programs prevent employees from misapplying information systems or planning to do so. Finally, [19] demonstrated that CSAT programs enhance employee cyber-security policy compliance behaviours. In the Irish context, the government of Ireland pointed out to the importance of the human factor and advised organisations to invest in employee education and awareness programs [20].

C. Cyber-Security in Ireland

The Irish government received a significant cyber-security breach in May 2021 when its health care services encountered

a major ransomware attack that largely damaged its processes for a long period of time [21]. Despite the remarkable disturbance caused by the attack, the Irish government rejected the hackers' demands. Irish cybersecurity agents collaborating with external partners executed countermeasures to disorder the hackers' activities. Ireland then introduced security measures and employed cybersecurity teams such as the Cyber Security Incident Response Team and the National Cyber Security Centre (NCSC) to reinforce Ireland's cyber-space through providing awareness, training, and consulting for Irish businesses [22]. Similarity in the financial sector, Bank of Ireland was charged with €1.66 million fine after one of its agents mistakenly transferred over €100.000 to a hacker who breached into a customer's email account a few years before the incident. Confidential information was released to the hacker over the phone without asking security questions or contacting the customer to double check his identity [23]. To answer the first research question on CSAT programs adoption in Ireland, we must classify the different financial institutions and their CSAT programs adoption. Please see Table 1.

Irish Financial Institutions	International Financial Institutions	Number of Employees	CSAT Adopted Y/N
Central Bank of Ireland		9,211 [24]	Y
Allied Irish Bank (AIB)		9,520 [25]	Y
The Irish League of Credit Unions		3,500 [26]	Y
Ulster Ireland Bank		> 3000 [27]	Y
Permanent TSB		2,400	Y
	KBC Bank Ireland	+1,000	N
EBS Ireland		360 (2018)	Y
	Citibank Ireland	2,500	N
	Danske Bank Ireland	> 650 [28]	N

Table 1: Financial Institutions Adoption CSAT Programs and Number of their Employees

D. Legislation relevent to cybercrime in Ireland

The recent Ireland's National Cyber Security Strategy (2019-2024) introduced key aspects linked to laws and legislations to comply with the EU in terms of reporting and cooperating [29]. Despite the limitations of the Irish criminal justice, it is very likely that Ireland now has enhanced its law in relation to cybercrime and that the powers of investigation for An Garda Siochana have been improver due to the criminal justice act 2011 that require financial organisations to report all sorts of fraud or online embezzlement to An Garda Siochana [30]. Such laws and legislations can positively persuade humans to comply to government policies [31].

D. Characteristics of an Effective CSAT Program

Cyber-security awareness training (CSAT) programs are adopted to improve employees' awareness, education and influence towards their organisation's information systems [32]. Effective CSAT programs must involve knowledge accessibility, knowledge retention, and knowledge sharing factors. Moreover, training must be based upon techniques that employees are aware of and constant practice to meet the desired training outcomes [33]. CSAT programs as a non-technical answer to cyber-threats but they must also highlight technical skills such as anti-phishing and password management tools, additionally, the effectiveness of CSAT program depends primary on the process and approach by which the program was conducted [34]. It depends on organisational as well as psychological factors such as support, memory, personality traits, beliefs, perceptions, and rewards [35]

Today, organisations need to conduct CSAT programs through effective management across all departments and functions. For instance, the marketing department might find CSAT program complex than the IT department. Therefore, management must design CSAT programs based on employees' skills, experience and knowledge to avoid gaps in CSAT programs which leads to ineffectualness.

E. Technology Acceptance Model (TAM)

The technology Acceptance Model (TAM) was first adopted by Davis in 1989 [36] as a framework to assume the likelihood of new technologies being used within organisations. It is so far one of the most discussed models among scholars in today's literature [37]. Please see Figure 2. TAM suggests that effective technology adoption is determined by the usage intention, which in return affected by the users' attitude towards the technology and its usefulness.

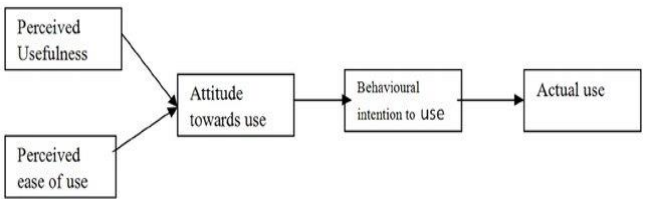


Figure 2: Technology Acceptance Model [36]

F. Unified Theory of Acceptance and Use of Technology (UTAUT)

The Unified Theory of Acceptance and Use of Technology (UTAUT) was developed by Venkatesh et al. in 2003 to unify eight information technology use and acceptance theories [38]. Please see Figure 3. Its aim is to justify both the employees and the organisation's behaviour in terms of their technology usage [39].

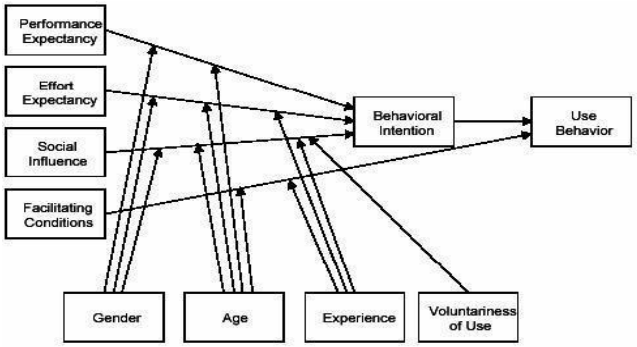


Figure 3 Unified Theory of Acceptance and Use of Technology (UTAUT) [38]

In terms of CSAT programs compliance, both TAM and UTAUT play a major role in CSAT programs effectiveness, with factors like social influence [36], experience and social attitudes [38], employees are more likely to comply and CSAT programs provided by their organisations.

CONCLUSION

This paper has demonstrated the importance of CSAT programs for Irish financial institutions. The findings indicate that all Irish-owned financial institutions conduct CSAT programs for employees as shown in Table 1. However, foreign-owned banks are noted as "No" is due either to the fact that information is unavailable online, or they do not adopt such programs, however this is not to say these programs are conducted internally. Therefore, further research is required to understand the level of employee's adoption and usage of CSAT and its overall benefits for organisations. Considering that research still indicates that human error remains the primary cause of cyber-attack. Moreover, this paper demonstrates that human factors are vital for CSAT program compliance and effectiveness. Additionally, laws and legislations play a critical role in employee cyber-security policy compliance. Finally, we advise a reconsideration in terms of CSAT program design based on employees' knowledge, memory, experience, and skills.

REFERENCES

[1] An Garda Síochána, "Cyber Crime ," *An Garda Síochána*, 2022.

[2] RTE, "Ireland sees biggest rise in cybersecurity attacks - CWSI survey," *RTE*, Jul. 26, 2021.

[3] MetaCompliance, " Security Awareness Training for the Financial Industry," *MetaCompliance* , 2021.

[4] E. U. Opara and O. J. Dieli, "Enterprise Cyber Security Challenges to Medium and Large Firms: An Analysis," *I.J. of Electronics and Information Engineering*, vol. 13, no. 2, pp. 77–85, 2021, doi: 10.6636/IJEIE.202106.

[5] A. Y. El-Bably, "Overview of the Impact of Human Error on Cybersecurity based on ISO/IEC 27001 Information Security Management," *Journal of Information Security and Cybercrimes Research*, vol. 4, no. 1, pp. 95–102, Jun. 2021, doi: 10.26735/wlpw6121.

[6] Z. (Justin) Zhang, W. He, W. Li, and M. Abdous, "Cybersecurity awareness training programs: a cost–benefit analysis framework," *Industrial Management and Data Systems*, vol. 121, no. 3, pp. 613–636, Mar. 2021, doi: 10.1108/IMDS-08-2020-0462.

[7] M. Zwilling, G. Klien, D. Lesjak, Ł. Wiechetek, F. Cetin, and H. N. Basim, "Cyber Security Awareness, Knowledge and Behavior: A Comparative Study," *Journal of Computer Information Systems*, vol. 62, no. 1, pp. 82–97, 2022, doi: 10.1080/08874417.2020.1712269.

[8] NCSC, "The National Cyber Security Centre (NCSC)," *NCSC*, 2021.

[9] S. Hu, C. Hsu, and Z. Zhou, "Security Education, Training, and Awareness Programs: Literature Review," *Journal of Computer Information Systems*, 2021, doi: 10.1080/08874417.2021.1913671.

[10] T. Mohammad, N. A. Mohamed Hussin, and M. H. Husin, "Online safety awareness and human factors: An application of the theory of human ecology," *Technology in Society*, vol. 68, Feb. 2022, doi: 10.1016/j.techsoc.2021.101823.

[11] M. I. Al-Ghamdi, "Effects of knowledge of cyber security on prevention of attacks," *Materials Today: Proceedings*, Apr. 2021, doi: 10.1016/j.matpr.2021.04.098.

[12] Central Bank of Ireland, "IT Risk in Credit Unions-Thematic Review Findings," Dublin, 2018. Accessed: Jan. 25, 2022. [Online]. Available: https://www.centralbank.ie/docs/default-source/Regulation/industry-market-sectors/credit-unions/communications/reports/it-risk-in-credit-unions-thematic-review-findings.pdf?sfvrsn=4

[13] A. Koohang, A. Nowak, J. Paliszkiewicz, and J. H. Nord, "Information Security Policy Compliance: Leadership, Trust, Role Values, and Awareness," *Journal of Computer Information Systems*, vol. 60, no. 1, pp. 1–8, Jan. 2020, doi: 10.1080/08874417.2019.1668738.

[14] E. Dincelli and I. S. Chengalur-Smith, "Choose your own training adventure: designing a gamified SETA artefact for improving information security and privacy through interactive storytelling," *European Journal of Information Systems*, vol. 29, no. 6, pp. 669–687, 2020, doi: 10.1080/0960085X.2020.1797546.

[15] M. Silic and P. B. Lowry, "Using Design-Science Based Gamification to Improve Organizational Security Training and Compliance," *Journal of Management Information Systems*, vol. 37, no. 1, pp. 129–161, Jan. 2020, doi: 10.1080/07421222.2019.1705512.

[16] M. Alqahtani and R. Braun, "Examining the Impact of Technical Controls, Accountability and Monitoring towards Cyber Security Compliance in E-government Organisations," 2021, doi: 10.21203/rs.3.rs-196216/v1.

[17] I. AL-ALAWI and S. AL-BASSAM, "The Significance of Cybersecurity System in Helping Managing Risk in Banking and Financial Sector," *Journal of Xidian University*, vol. 14, no. 7, Jul. 2020, doi: 10.37896/jxu14.7/174.

[18] K. L. Gwebu, J. Wang, and M. Y. Hu, "Information security policy noncompliance: An integrative social influence model," *Information Systems Journal*, vol. 30, no. 2, pp. 220–269, Mar. 2020, doi: 10.1111/isj.12257.

[19] E. Kweon, H. Lee, S. Chai, and K. Yoo, "The Utility of Information Security Training and Education on Cybersecurity Incidents: An empirical evidence," *Information Systems Frontiers*, vol. 23, no. 2, pp. 361–373, Apr. 2021, doi: 10.1007/s10796-019-09977-z.

[20] Government of Ireland, "12 Steps to Cyber Security Guidance on Cyber Security for Irish Business," Dublin, Oct. 2018. Accessed: Mar. 31, 2022. [Online]. Available: https://www.ncsc.gov.ie/pdfs/Cybersecurity_12_steps.pdf

[21] HSE, "Conti cyber attack on the HSE Independent Post Incident Review Commissioned by the HSE Board in conjunction with the CEO and Executive Management Team," 2021. Accessed: Jan. 26, 2022. [Online]. Available: https://www.hse.ie/eng/services/publications/conti-cyber-attack-on-the-hse-full-report.pdf

[22] C. Keena, "Opening of email attachment led to HSE cyber attack, report finds," *The Irish Times* , Dec. 10, 2021. https://www.irishtimes.com/news/crime-and-law/opening-of-email-attachment-led-to-hse-cyber-attack-report-finds-1.4752043 (accessed Jan. 27, 2022).

[23] L. Boland, "Bank of Ireland fined €1.6 million by Central Bank following cyberfraud investigation," *Journal.ie*, Jul. 28, 2020. https://www.thejournal.ie/bank-of-ireland-cyberfraud-fine-investigation-5161825-Jul2020/ (accessed Jan. 28, 2022).

[24] Bank of Ireland, "Craft," *Bank of Ireland* , 2022. https://craft.co/bank-of-ireland (accessed Jan. 26, 2022).

[25] RTE, "AIB to cut 1,500 jobs by 2022 as profits drop," *RTE*, Mar. 06, 2020.

[26] Gov.ie, "Credit Unions and the Irish Economy," *Gov.ie* , 2019. https://www.gov.ie/en/publication/26d557-credit-unions-and-the-irish-economy/ (accessed Jan. 26, 2022).

[27] Ulster Bank, "About Us," *About Us* , 2022. https://www.ulsterbank.ie/globals/about-us/corporate-information.html (accessed Jan. 26, 2022).

[28] European Banking Resources, "National Irish Bank ," *National Irish Bank* , 2022. https://www.ecbs.org/banks/ireland/national-irish-bank/view-details.html#:~:text=With%20over%20650%20employees%20operating,unique%20SME%20Business%20Banking%20services. (accessed Jan. 26, 2022).

[29] S. Brady and C. Heinl, "Cybercrime: Current Threats and Responses A review of the research literature," 2020. Accessed: Mar. 31, 2022. [Online]. Available: https://www.justice.ie/en/JELR/Cybercrime_-_Current_Threats_and_Responses.pdf/Files/Cybercrime_-_Current_Threats_and_Responses.pdf

[30] Cyber Crime Ireland, "Irish Laws ," *Cyber Crime Ireland*, 2022.

[31] C. Friend, L. B. Grieve, J. Kavanagh, and M. Palace, "Fighting Cybercrime: A Review of the Irish Experience," *International Journal of Cyber Criminology*, vol. 14, no. 2, pp. 383–399, 2020, doi: 10.5281/zenodo.4766528.

[32] T. Mashiane and E. Kritzinger, "IDENTIFYING BEHAVIORAL CONSTRUCTS IN RELATION TO USER CYBERSECURITY BEHAVIOR," *EURASIAN JOURNAL OF SOCIAL SCIENCES*, vol. 9, no. 2, pp. 98–122, 2021, doi: 10.15604/ejss.2021.09.02.004.

[33] I. Legárd, "EFFECTIVE METHODS FOR SUCCESSFUL INFORMATION SECURITY AWARENESS," *Pro Publico Bono - Magyar Közigazgatás*, vol. 9, no. 1, pp. 108–127, Aug. 2021, doi: 10.32575/ppb.2021.1.7.

[34] M. Alshaikh, "Developing cybersecurity culture to influence employee behavior: A practice perspective," *Computers & Security*, vol. 98, p. 102003, 2020, doi: 10.1016/j.cose.2020.102003.

[35] L. Hadlington, "Human factors in cybersecurity; examining the link between [3 _ T D $ D I F F] Internet addiction, impulsivity, attitudes towards cybersecurity, and risky cybersecurity behaviours," *Heliyon*, vol. 3, p. 346, 2017, doi: 10.1016/j.heliyon.2017.

[36] F. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989, Accessed: Mar. 31, 2022. [Online]. Available: https://doi.org/10.2307/249008

[37] H. Guner and C. Acarturk, "The use and acceptance of ICT by senior citizens: a comparison of technology acceptance model (TAM) for elderly and young adults," *Universal Access in the Information Society*, vol. 19, no. 2, pp. 311–330, Jun. 2020, doi: 10.1007/s10209-018-0642-4.

[38] V. Venkatesh, R. H. Smith, M. G. Morris, G. B. Davis, F. D. Davis, and S. M. Walton, "USER ACCEPTANCE OF INFORMATION TECHNOLOGY: TOWARD A UNIFIED VIEW," *MIS Quarterly* , vol. 27, no. 3, pp. 425–478, 2003, Accessed: Mar. 31, 2022. [Online]. Available: https://doi.org/10.2307/30036540

[39] A. Ayaz and M. Yanartaş, "An analysis on the unified theory of acceptance and use of technology theory (UTAUT): Acceptance of electronic document management system (EDMS)," *Computers in Human Behavior Reports*, vol. 2, p. 100032, Aug. 2020, doi: 10.1016/j.chbr.2020.100032.

A contribution towards the regulation of anonymised datasets within the framework of GDPR

F. Cormac Britton
Dept. of Computing
ATU Mayo
cormacbritton1991@gmail.com

Seamus Dowling
Dept. of Computing
ATU Mayo
Seamus.Dowling@gmit.ie

Mark Frain
Dept. of Computing
ATU Mayo
Mark.Frain@gmit.ie

***Abstract*—The European Union's General Data Protection Regulation legal framework outlines data anonymisation as an effective method to help data processors meet their legal obligations. Once personal data has been anonymised, GDPR no longer recognises it as personal data and it becomes free to be processed unregulated. The potential capacity for re-identification of any anonymised dataset brings into question the value of anonymisation and whether it should be regulated to ensure the privacy of data subjects. This paper identifies and reviews some of these problems, challenges the current set of regulations, and recommends maintaining regulation over anonymised datasets, and enforcing systems of traceability and erasure of datasets over time, which would increase the value of utilising anonymisation and its capacity to be a reliable system for data privacy.**

***Index Terms*—personal data, anonymisation, GDPR, re-identification**

I. Introduction

Under the legal framework of the European Union's General Data Protection Regulation (GDPR), the legislation that deals with matters pertaining to the processing of personal data in the EU, data controllers and data processors are obliged to follow the regulations when processing personal data (any form of data that can be used to identify a natural person). Before personal data can be processed, consent from the data subject (the individual who can be identified from the data) must first be legally obtained [1]. As long as the data processor is processing personal data in the EU in any way, it falls under the regulations of GDPR.

Data anonymisation is a sanitation method for removing identifiable information from datasets containing personal information or altering that information in such a way that it is no longer considered to be identifiable. Once data has been anonymised, it is no longer considered to be personal data and therefore is no longer under the scope of GDPR and not under any regulation or protection [2]. However, the process of anonymisation is still data processing and consent should still be obtained first.

The utilisation of anonymisation itself can be considered a useful means of risk reduction and a means of GDPR compliancy for the data controller and processor [3]. Employing anonymisation to sanitise personal information as the default method could support particular data management strategies like privacy by design or data minimisation. Most anonymisation techniques don't require technical knowledge and could potentially prove to be an effective and efficient method to help data processors comply with their GDPR obligations.

Funded by ATU Mayo.

There are some areas of concern regarding the employment of anonymisation. Several case studies have shown that even after anonymisation, personal data isn't immune from re-identification [3]. Some re-identification techniques show certain weaknesses in anonymised datasets and how they can be exploited. It should be made clear, however, that re-identification is not a reversal of the anonymisation techniques used to anonymise the data initially. Anonymisation technique are irreversible and once a dataset has been anonymised, the original data cannot be directly obtained from the anonymised data. Rather, re-identification techniques depend on multiple sets of anonymised data to make links by inference.

The general goal of anonymisation is not to make the data completely illegible such that it can no longer be meaningfully processed. Rather, the goal is to remove the link between the data and the data subject such that the data subject cam no longer be identified. In the final state, a lot of data will still be legible, depending on the needs of the data processor to retain some informational value so that it is still useful, e.g. for studies, metrics, decision making, marketing, etc. Generally speaking, it is the dataset that has been anonymised, severing a direct link to the data subjects, but it could be argued that this does not guarantee anonymisation in general as combining multiple sets of anonymised data can re-establish the link to the data subjects.

It is understood that perfect anonymisation is effectively impossible to achieve to protect against potential re-identification. As long as the anonymised dataset contains meaningful information that can indirectly link to a data subject, it poses an inherent risk [4]. While individual sets of anonymised data are considered truly anonymised so that they fall beyond the scope of GDPR [2], without any form of protection against re-identification, the risk with handling anonymised data has the potential to grow significantly in the long term.

There are no metrics for judging the effectiveness of applied anonymisation. The legislation gives the power to the data controller to determine that the applied anonymisation techniques are sufficient. The legislation itself does not outline

the techniques that should be employed or any means to determine the effectiveness of any technique that might be employed. There exist a number of guidelines that do outline such techniques and their general effectiveness [5]. Though, it is determined that there is no way to quantify anonymisation effectiveness and ultimately it will be human judgement that determines whether anonymisation has been achieved [2].

As anonymisation cannot truly be quantified, as the standards are vague and left to human interpretation and judgement, and as anonymised datasets pose an inherent risk, it is hard to argue that most anonymised data sets are truly anonymous (at least, as long as they contain any legible information that could indirectly link to a data subject). Therefore, it should be considered that the structures for anonymisation, as they currently are, aren't fit for protecting personal information.

However, anonymisation itself, whether it is serves well or not, is defended as a legitimate technique in the current legislation [1]. Indeed, there is much potential for misuse of anonymised datasets for malicious purposes, personal gain or otherwise, and with little to safeguard against such misuse, those who would abuse anonymised data might not have to fear any consequences. This paper aims to analyse the issues around anonymisation and put forth contributions accordingly.

II. Methodology

The methodology incorporated into this research will primarily be of a qualitative approach through a review of anonymization literature, re-identification literature and legal literature and relevant case studies, interpreting the data to either support the goals of the research or to be analysed further alongside other data.

Relevant case studies will be reviewed, providing important insight into scenarios of anonymisation when put into practice, the re-identification of anonymised data, the implications of unauthorised identification, and the reflection such implications have on the current legal framework in GDPR.

It will also include cross referencing multiple sources, such as the GDPR legal framework, official guidelines on how to become GDPR compliant, research papers, research journals, and case studies, all in respect to data anonymisation, re-identification techniques, and legal liability in respect to these areas. These methods, while taking a qualitative approach in content analysis, will also help quantify data to compare with criteria and other analysed data, with an aim to answer each research question.

III. Previous Research and Analysis

A. Anonymisation

Anonymous information itself is defined as information which does "does not relate to an identified or identifiable natural person" or as information where the "data subject is not or no longer identifiable" [1]. Once the data has been anonymised, it is "irreversibly preventing the identification of the individual" [4]. This anonymised data is then considered "no longer personal data and data protection legislation no longer applies [2]." The data controller is to use "all the means reasonably likely" to ensure that the data has effectively undergone anonymisation [1]. The data controller is to judge that "singling out, linkability and inference" has been sufficiently minimised during the anonymisation process [4].

When considering the effectiveness of anonymisation, consideration must be taken of the attributes which can be linked from multiple datasets, known as *quasi-identifiers*, where such data gets released into the world as microdata (e.g. medical records, voter registration, etc.) and where such data can be used for research or public benefit [6]. While such microdata can be used for beneficial research or benign interest, there is also a trade-off with general data privacy.

An interesting area in anonymous data is the area in data mining. *k*-anonymity is defined as a property of sets of information where there is a measure of protection against identifying the personal data of an individual in a given record [6]. While *k*-anonymous datasets are useful for data-mining, it has been shown that *k*-anonymity itself does not guarantee privacy [7]. It has been proposed that "trivial" sanitation methods be used by separating "quasi-identifiers from sensitive attributes" rather than using generalisation and suppression techniques on *quasi-identifiers* to prevent making the data-mining utility useless [7]. However, most anonymisation algorithms utilise generalisation and suppression techniques.

The main anonymisation techniques are *randomisation*, *generalisation* and *masking*. *Randomisation* is the altering of data by adding "noise" or changing the data itself. For example, making small changes to the heights of individuals while stating clearly how the data is accurate within a range of values. Within *randomisation*, *permutation* involves swapping certain records where the data needs to be accurate but the correlation doesn't need to be maintained. *Generalisation*, by the process of *k*-anonymity, involves ensuring that enough data subjects fall within a certain band that can be generalised such that no particular data subject stands out. *Masking* is intended to be applied on top of other anonymisation techniques to improve anonymisation (as a standalone anonymisation technique, *masking* would be ineffective) removing all obvious identifiers such as names and addresses [4], [8]. In practice, anonymisation techniques include the redaction of names, blurring faces in video footage, disguising identifying features in audio material and altering details in reports using practices such as generalisation. The process is considered time consuming and relying on careful human judgement to ensure that data is sufficiently anonymised [5].

Another technique involves "secure-keyed cryptographic hash" functions, deleting the keys once the data has been hashed (the data would only be pseudonymised if the keys weren't deleted, and still considered to be personal data protected under GDPR) [5]. Such hashing techniques can be used in the context of anonymisation so long that all information has been removed that allows for re-identification and the system is judged to be robust against re-identification attacks [9]. In order to be able to make effective judgements on the sufficiency of how well data is anonymised, it is recommended that the data controllers perform a "motivated intruder test"

by applying rigorous testing methods which often involves studying the data in secondary sources that an intruder might have access to in order to compare to data to check for linkable information that might lead to re-identification. These secondary sources are broad and may include social media, newspapers, genealogy websites and libraries [5].

B. Re-identification

One effective method of re-identifying individuals from anonymised datasets is known as the *inference attack* where correlated events and information from multiple datasets can be used to re-identify a data subject and to learn more private information about that subject. One study looks at re-identification techniques based on matching geolocation data of individuals from different datasets, emphasising that only using pseudonyms is not a sufficient means of maintaining the integrity of anonymised personal data. More specifically, with the amount of personal phones equipped with GPS systems, and with the amount of data processing that goes into the collected GPS data from individuals, the potential impact of mass re-identification based on geolocation data could be lead to compromising the private information of an extremely large number of data subjects [10]. In another method, anonymised data can still be singled out to identify a data subject [8].

One case study looks at microdata, "such as individual preferences, recommendations, transaction records", etc., as a means of re-identifying datasets through the *inference attack* where the Netflix records of its subscribers were compared with movie reviews on the Internet Movie Database, not only identifying the individual data subjects but also their political affiliations [11]. In a follow-up study, it is shown that identifying individual data subjects through microdata inferences is not a new concept, but rather "the core technical insight goes back at least 60 years" and continues to make the point that "high-dimensional data is inherently vulnerable to de-anonymization", as the research supports itself with not only theoretical evidence, but with "robust de-anonymization techniques" applied to data sources such as geolocation data, credit card data, browsing data, and even source code and binary files [12].

One study explores how machine learning systems can re-identify text data through an *inference attack* based on a data subject's texting habits and usage of predicative text compared with text data found elsewhere, such as emails, social media posts, forums and blogs. By exploring settings where users wish to remain anonymous, specifically the infamous Silk Road online black market, a data subject could be re-identified by analysing their text "fingerprint" [13]. Another study deals with re-identifying social media data by comparing data on different social media platforms, making the argument that anonymisation itself does not guarantee any privacy in the context of social media [14]. A third study shows that even human mobility is unique to the individual where even just studying four spatio-temporal data points would be enough to correctly identify 95% of individuals [15]. A follow-up study shows that credit card metadata between four spatio-temporal data points is enough to correctly identify 90% of individuals [16].

As technology rapidly evolves, as devices are increasingly being connected and as individuals all across the world are more likely to have access to a personal device, everyone puts themselves at risk of having one's right to privacy compromised. Not only that, but general data processing seems to inherently contain risk of identifying the private data of individuals. Even heavily sampled anonymised datasets can lead to the re-identification of the vast majority of individual data subjects [17]. It is also possible to identify individuals from publicly available census summaries, suggesting that this type of data is not in fact anonymous, pointing out that that publicly available health data also fall under this issue [18].

The European Union Agency for Cybersecurity (ENISA) outlines the types of attacks that can occur on sets of pseudonymised data: the brute force attack through computational methods or access to "black box" implementations of the pseudonymisation technique; dictionary attacks by pre-computing a large number of pseudonyms and saving them into a dictionary in an attempt to identify data in the pseudonymised dataset with the same values by comparison; making educated guesses by utilising background knowledge that may be related to the pseudonymised data. While pseudonymised data is not anonymised data, data can be anonymised by undergoing cryptographic pseudonymisation and then, by deleting the generated keys, the data can be effectively anonymised [19].

C. The legal framework

Where it is argued that pseudonymised data be considered as anonymised data rather than personal data in order to improve the freedoms for data processors, it is suggested that processes for data governance systems such that policies can be put into place to safeguard such data through contractual terms of use for data processors specifying that "the data must be used for research purposes only and the researcher may make no attempt to re-identify any individuals within the data; and a policy which sets out the penalty for any breach of the terms of use" [20].

Consider the legal consequences and liabilities an organisation may or should find itself under if anonymised data that it had made available resulted in the re-identification of private personal data of individual data subjects. If an organisation releases a dataset of anonymised data for processing purposes, given the legal protections that the data is no longer considered to be personal data, they are in effect taking the risk of allowing that data to be potentially re-identified in the future. While an organisation can be investigated to ensure that they took the proper precautions and followed their own security policies, it may not be enough. In fact, as pointed out in the conclusion, rigorous re-identification testing of anonymised datasets may only ensure that there is an increased challenge for a third party to re-identify a dataset rather than outright prevention [21]. Based on this, it can be argued that anonymised data requires its own regulations in order to protect the privacy

rights of the individual and in order to mitigate the damage done once a anonymised dataset has been re-identified.

As one study points out, even heavily sampled datasets of anonymised datasets can lead to the re-identification of 99.98% of data subjects, challenging the lack of regulation in GDPR towards anonymised datasets and subsequently the lack of liability of the data processors who are free to process and release anonymised datasets without fear of consequence [17]. Another suggests that in this environment that the "burden of proof be on the data controller to affirmatively show that anonymized data cannot be linked to individuals, rather than on privacy advocates to show that linkage is possible" [12].

Consider the privacy uncertainties involved in the processing of anonymised data as advancements are already being made in re-identification techniques while data, anonymised or otherwise, becomes more ubiquitous. An argument can be made that such uncertainties influence entities that process the data as downplaying any potential privacy risks when handling anonymised data. A study identifies and examines the role of anonymised data under GDPR as from the perspective of the processors. The roles as pointed out by the article include utilising anonymisation techniques as a measure to avoid privacy rules for data processing; to avoid certain data privacy obligations, including the obligation to report data breaches; and as a general method for data privacy compliance. These roles could be considered more of a reflection of convenience for data processors for the purposes of data privacy regulations avoidance rather than as a tool to be used with care only when necessary [22].

Another area of concern when it comes to data privacy is big data. D'Acquisto et al. highlight the concerns of maintaining data privacy as big data processing is rapidly expanding. One concern is the lack of control by the data processor of big data, being unaware of what data is being processed and all of its sources and how it flows between systems. Another concern is data re-usability where how one uses websites and apps on their personal devices leads to the personal habits being processed in agreed circumstances but also being valued by third party entities. Similarly, the data profiling of individuals may be designed for automated marketing or convenience purposes but can also be used in discriminatory practices. Strategies to promote privacy by design in the big data context are highlighted to include minimising collected data, aggregating collected data, separating data in storage, and enforcement through policy [23]. The Privacy Preserving Techniques Task Team (PPTTT) and the UN Global Working Group (GWG) apply more specific focus in big data where they propose a number of privacy base policy and encryption methods, looking at example usages with use cases, adversarial perspectives, security arguments and costs of usage under a privacy by design focused framework [24].

Regarding the rights of the data subject, one study differentiating non-personal data from personal data from a "law and computer science perspective". It makes the argument that it is important to be able to make such a distinction in order to understand the scope of regulatory application and also explores the idea and difficulties of determining de-personalised data as non-personal data, in that where the data was once personal has since undergone anonymisation techniques rendering it non-personal. This article also acknowledges the legal ambiguities regarding the definition of anonymised data. It points out the GDPR itself is open to levels of risk of identification while others would insist that any risk is unacceptable. Through reviewing the technicalities of anonymisation with case studies, it concludes that "there always remains a residual risk when anonymisation is used". This outlines that not only are the defined boundaries of non-personal data ambiguous within the legal framework, it is also clear that the legal framework lacks a strength in its protection of personal data that could be identified from unprotected anonymised data that always contains some level of risk [3].

IV. Findings

A. Preventing re-identification

It is not generally possible to determine how likely of a risk that an anonymised dataset can pose in the re-identification of personal data [4]. The risk of inference from another anonymised dataset varies on the amount of existing datasets of related personal data that has been anonymised, and also on the degree to which each anonymised dataset is sanitised and the capacity for each anonymised dataset to contain data that is inferable with other anonymised datasets.

Depending on the nature of the anonymised dataset itself, regarding its content and the relationship such content has with personal information, even if personal information cannot be directly deduced from the dataset, it is effectively impossible to prevent an inference attack if the dataset is accessible. Any anonymised datasets that become published are at risk of inference attacks when combined with other existing datasets or even datasets that are yet to exist. Precaution would have to be taken to avoid publishing anonymised datasets, but this would render such a dataset useless if publication is the goal for research purposes or otherwise. Any anonymised datasets that aren't intended to be published could be kept from doing so in order to prevent any potential re-identification attacks.

Other re-identification techniques, such as re-identification through predictive patterns, are likely extremely difficult, if not impossible to avoid. Techniques that exploit social media and other social interactions could be avoidable if one hides themself away from the world and not engage with technology, but this doesn't seem like an effective solution either.

While some steps can be taken to reduce the exposure to re-identification attacks, it is not fully possible nor practical to avoid these attacks completely.

B. The capacity for misuse of anonymised datasets

There is always an inherent risk of re-identification with anonymised datasets. Over time, one could expect this risk to increase for multiple reasons:

- without the regulation of anonymised data, data processors can handle anonymised datasets in any manner they

wish, increasing the likelihood of misuse in cases of re-identification for personal gain;
- without the regulation of anonymised data, anonymised datasets could exist indefinitely and pass many hands, increasing the likelihood that re-identification will occur within a matter of time and if the anonymised dataset gets into the wrong hands;
- with a lack of clarity over the legality of the re-identification of personal data, there could be an incentive to misuse anonymised datasets in cases of re-identification without repercussion;
- and when more and more anonymised datasets become available, the risk inferring information between two or more anonymised datasets significantly increases.

As long as these anonymised datasets continue to exist indefinitely, their capacity for misuse would likely become more and more probable.

C. Reviewing the legal framework

Where it comes to the legal framework of GDPR, there is much room given for data processors to utilise processes such as anonymisation in order to process data freely. Once the data has been anonymised, the data processor is no longer under any data protection regulations for that data. In the case that anonymised data becomes re-identified by an unauthorised person, there does not appear to be any liability on the data processor or controller.

The data subject, who's data is undergoing the processing, is only protected so long as the data is not considered to be anonymised. Once the data has been anonymised, the regulations and protections of GDPR no longer apply to the data subject. Of consequence, if an anonymised data set is utilised in re-identification of the subject, then the subject bears all of the consequence of having their personal data exposed. Without liability over the processing of anonymised data, the data subject has no control over this data and lacks any recompense in the case of re-identification.

Effectively, it is clear that when data is anonymised, and no longer considered personal information, that it loses all of the protections and regulations of personal information. This can have consequences on data subjects when it comes to their data privacy. It can also have consequences of entities motivated by personal gain to manipulate data in ways that aren't protected by the law but could have a large negative impact on a significant number of data subjects.

V. Recommendations

As it is not generally possible to grade the quality of anonymised data against possible re-identification with a metric, a system based on responsibility and accountability is justifiable in respect to the measures taken to ensure that the personalised data has been rigorously anonymised. Further, this paper suggests that anonymised data itself be regulated in its own special category of data (as long as anonymised data is to remain distinct from personal data), such that it is recognised that all anonymised data has the capacity to be utilised in the re-identification of personal data. Without enforceable regulations that ensure the protection of personal data from being re-identified through the utilisation of anonymised datasets, personal data is not adequately protected in that it is in danger of being exploited by a function of GDPR itself: anonymisation.

Possible measures to overcome the lack of feasible graduated metrics for anonymisation processes are systems of:

- traceability, such that anonymised data sets can be linked to the processes behind the anonymisation, including purposes for anonymisation, specific anonymisation techniques that were employed, details regarding what final appearance the anonymised data should take for the goals of further processing, and records of transaction with other parties;
- lifespan, such that all anonymised datasets have a set lifespan, in accordance with any potential future regulations, and that anonymised datasets should be erased by the end of this lifespan;
- and accountability, such that the data controller took the necessary steps to ensure that the employed anonymisation techniques were rigorous enough such that the final state of the anonymised dataset is justifiable in respect to the purposes and goals for further processing, and that the anonymised dataset was erased and underwent no further processing since the end of its lifespan.

In order to ensure traceability, anonymised datasets could be assigned a unique ID specific to the data controller or processor, but arbitrary in respect to the content of the data itself. The unique ID should not reveal nor contribute to the re-identification of personal data. In respect to the lifespan, the time-to-expire could be embedded into the unique ID or recorded separately. The Data Controller should ensure that there are processes in place to ensure that the anonymised dataset undergoes no further processing and is completely erased by the time-to-expire. Software systems could be put into place for the handling of anonymised datasets and their lifespans, ensuring automatic erasure at the time-to-expire, and for maintaining records of traceability.

One issue that needs to be further considered is how to handle the transaction of anonymised datasets between parties regarding the traceability, lifespan and accountability. Potentially, a new unique ID could be generated by the third party receiving the anonymised dataset with records of the transaction and a new time-to-expire. Another possibility is to retain the original unique ID and time-to-expire, but modifying the unique ID or record to include information about the transaction and new ownership.

Any policies involving the anonymisation of personal data and the handling of anonymised datasets should outline:

- sanitation techniques to reduce the residual risk of misuse for re-identification;
- that the goals for the processing of the anonymised dataset are to be marked clearly;

- that enough sanitation techniques are to be applied to reduce this risk to a minimum while retaining a deliverable state for further processing within the outlined goals;
- that the applied sanitation techniques are to be recorded for traceability;
- that a unique ID should be applied to the final anonymised dataset for traceability;
- that an appropriate time-of-expiry should be determined for the anonymised dataset;
- that the data subjects be first consulted before the publishing of their personal data, even if that data is later anonymised;
- and that there are to be systems in place to ensure the erasure of the anonymised dataset within the time-of-expiry.

In order to ensure a system of accountability, data controllers and processors could include these matters as a part of their own data protection policies. Ensuring that anonymised datasets, the applied sanitation techniques and the time-of-expiry are identifiable, a system of accountability for the data controller and processor can be encouraged. However, being under no obligations to do so, there would be little incentive for organisations that process data to put these systems into place. A change in legislation would be required, enforcing measures to be taken, as a matter of legal obligation, to regulate the processing of anonymised datasets.

VI. Conclusion

Ultimately, this paper challenges the systems of GDPR to consider the outlined issues regarding data anonymisation and the likelihood of re-identification, to consider the impact these issues have on personal data due to the potentiality, or inevitability, of re-identification, to consider the means to mitigate against such potentialities, and to consider making the necessary legislation to enforce the mitigation necessary for the long-term protection of personal data.

In order to change or add legislation to GDPR, in respect to the regulation of anonymised data, anonymised data would need to be re-categorised as a form of personal data. Potentially, new legislation in conjunction to GDPR, but not under the scope of GDPR, could acknowledge anonymised data under its own special category with its own regulations. However, as a point of regulation, in respect to cases where personal data is re-identified by means of utilising anonymised datasets, it may be necessary to expand the scope of GDPR since the fundamental issue involves the re-identification (and subsequent unlawful processing) of personal data from these anonymised datasets.

References

[1] "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," European Union, Official Journal of the European Union, L 119, vol. 59, May 2016.

[2] Handbook on European data protection law, the European Union Agency for Fundamental Rights (FRA), the European Court of Human Rights (ECtHR), the Council of Europe (CoE) and the European Data Protection Supervisor (EDPS), April 2018.

[3] M. Finck and F. Pallas, "They who must not be identified—distinguishing personal from non-personal data under the GDPR," in International Data Privacy Law, vol. 10, no. 1, pp. 11-36, March 2020.

[4] Guidance Note: Guidance on Anonymisation and Pseudonymisation, An Coimisiún um Chosaint Sonraí (Data Protection Commission), June 2019.

[5] Anonymisation: managing data protection risk code of practice, Information Commissioner's Office, November 2012.

[6] G. Ghinita, P. Karras, P. Kalnis and N. Mamoulis, "Fast data anonymization with low information loss," in VLDB '07: Proceedings of the 33rd international conference on Very large data bases, VLDB Endowment, September 2007.

[7] J. Brickell and V. Shmatikov, "The cost of privacy: destruction of data-mining utility in anonymized data publishing," in The 14th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD, 2008.

[8] Opinion 05/2014 on Anonymisation Techniques, Article 29 Data Protection Working Party, April 2014.

[9] Introduction to the Hash Function as a Personal Data Pseudonymisation Technique, Agencia Española de Protección de Datos (AEPD, Spanish Data Protection Agency) and European Data Protection Supervisor (EDPS), October 2019.

[10] S. Gambs, M.-O. Killijian and M. N. del Prado Cortez, "De-anonymization attack on geolocated data," in Journal of Computer and System Sciences, vol. 80, April 2014.

[11] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in 2008 IEEE European Symposium on Security and Privacy, IEEE, May 2008.

[12] A. Narayanan and V. Shmatikov, Robust de-anonymization of large sparse datasets: a decade later, May 2019.

[13] Z. Sun, R. Schuster and V. Shmatikov, "De-anonymizing text by fingerprinting language generation," in Advances in Neural Information Processing Systems 33 (NeurIPS 2020), NeurIPS, 2020.

[14] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in 2009 IEEE European Symposium on Security and Privacy, IEEE, May 2009.

[15] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen and V. D. Blondel, "Unique in the crowd: the privacy bounds of human mobility," in Scientific Reports, vol. 3, March 2013.

[16] Y.-A. de Montjoye, L. Radaelli, V. K. Singh and A. Pentland, "Unique in the shopping mall: on the reidentifiability of credit card metadata," in Science, vol. 347, no. 6221, January 2015.

[17] L. Rocher, J. M. Hendrickx and Y.-A. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," in Nature Communications, vol. 10, July 2019.

[18] L. Sweeney, "Simple demographics often identify people uniquely," in Health, vol. 671, January 2000.

[19] Pseudonymisation Techniques and Best Practices, European Union Agency for Cybersecurity (ENISA), November 2019.

[20] M. Mourby, E. Mackey, M. Elliot, H. Gowans, S. E. Wallace, J. Bell, H. Smith, S. Aidinlis, J. Kaye, "Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK," in Computer Law & Security Review, vol. 34, no. 2, April 2018.

[21] C. C. Porter, "De-identified data and third party data mining: the risk of re-identification of personal information," in Washington Journal of Law, Technology & Arts, vol. 5, no. 1, September 2008.

[22] S. Y. Esayas, "The role of anonymisation and pseudonymisation under the EU data privacy rules: beyond the 'all or nothing' approach," in European Journal of Law and Technology, vol. 6, no. 2, October 2015.

[23] G. D'Acquisto, J. Domingo-Ferrer, P. Kikiras, V. Torra, Y.-A. de Montjoye and A. Bourka, "Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics," European Union Agency for Cybersecurity (ENISA), December 2015.

[24] UN Handbook on Privacy-Preserving Computation Techniques, Privacy Preserving Techniques Task Team (PPTTT) and UN Global Working Group (GWG), March 2019.

Beware of Titles: Analysing Media Reporting of Cybercrime in UK and UAE

Maitha Khaled Al Mazrouei
College of Technological Innovations
Zayed University
Abu Dhabi, UAE
M80008364@zu.ac.ae

Danica Čigoja Piper
College of Communication and Media Sciences
Zayed University
Abu Dhabi, UAE
danica.piper@zu.ac.ae

Lena Yuryna Connolly
College of Technological Innovations
Zayed University
Abu Dhabi, UAE
0000-0002-7110-9594

***Abstract*—In this paper, media reporting of cybercrime victims (with a specific focus on organisations) was examined. For this purpose, data were collected from media outlets in the UK and UAE. A basic premise of this paper is that media has the ability to restrict readers to either 'opposing' or 'favouring' views about the subject of the news. Research findings demonstrated that victim organisations were mostly shaped within a negative media frame in a given time period. Comparative analysis between UK and UAE news media showed the usage of similar language in portraying victim organisations.**

Keywords—media framing, cybercrime reporting, victim organisations, UK and UAE

I. Introduction

News media has three primary roles in a society – to inform, to educate and to entertain [1]. There is also a social responsibility factor that demands from the media to bring attention to the issues that are important to the public. These issues, however, can be framed in a certain way, inducing readers to form either 'opposing' or 'favouring' views about the subject of the news [2]. *Framing* is a theory of mass communication that elucidates how the media packages and presents information, subsequently shaping public opinion. The framing theory has been commonly utilised to investigate how the news media frames various important issues. Reference [3], for instance, conducted a study on media framing of COVID-19 in China and found out that the media used frames like 'war', 'race', 'chess' and 'challenge' to shape the ideology of individuals about the virus. Furthermore, reference [4] reported on how environmental activists made use of the COVID-19 pandemic by employing frames like 'humans are the biggest virus', 'against animal exploitation' and 'changing our lifestyle' to bring public's attention to the urgencies of climate change. Moreover, reference [5] investigated media coverage of fetal alcohol spectrum disorders (FASD) and discovered the dominant frames of 'sympathy' and 'shame'. Essentially, media encouraged the feeling of sympathy towards children with FASD, but at the same time, portrayed shame for 'deserving' mothers.

Among many important tech and security news topics, cybercrime is one such issue. Media reporting is actively involved into the process of forming a public opinion about cybercrime [6, 7]. Arguably, media has the ability to shape a public opinion about victims of cybercrime, portraying them with either positive or negative frames. A literature search on the topic, however, revealed limited research. Reference [6], for instance, inspected articles printed in 2011-2016 by two popular British tabloids (i.e., Daily Mirror and The Sun) and discovered that cybercrime was framed as a source of 'social danger' and 'fear'. Additionally, English and Russian media represented cyber-attacks via frames of 'war', 'game', 'pandemic' and 'crime' [8]. Moreover, reference [9] discovered that the framing of cybercrime in Nigeria is useful in eradicating it. Several studies reported that media framing has an impact on public perceptions about cybercrime and its victims [10, 11]. Although these works are important contributions to the pool of knowledge in the realm of cybercrime framing, no studies were conducted to specifically investigate news media reporting practices of cybercrime victims (organisations, in particular). This is important because victim organisations are unwilling to share information about cyber-attacks due to fear of negative publicity [7, 12]. This, in turn, serves as a discouragement to be forthcoming with the information about these attacks. Platforms for cybersecurity information sharing, however, have been acknowledged as an important defense against cybercrime [13]. News media can be viewed as one such information sharing platform due to its duty to bring attention to the issues important to the public. In light of the above, the objective of this study is to examine media coverage of cybercrime victims in two contrasting cultural environments, specifically in the UK and UAE (with the view to expand this study in other countries).

A specific focus of this work is on articles' headlines (also, interchangeably referred in this study as 'titles') as this part of the text can give researchers important insights about the angle that media takes regarding a given topic. Titles are considered as an important element in the news text analysis because this is the first piece of information that readers get from the text [14]. Furthermore, researchers highlighted that a significant number of newspaper readers only pay attention to titles in order to form their opinion about a specific topic and therefore overlook additional clarifications in articles' body text [15, 16].

II. Theoretical Framework

Framing theory was originated with the work of anthropologist Gregory Bateson in 1950's who introduced the idea 'framework as "as a tool of the psyche that explains why people focus their attention of some stylized aspects of reality and not others" [17, p.1]. Although the framing theory gained popularity in various disciplines [18], it has received the utmost consideration in media and communication science. According to reference [19], in communication research *framing theory* postulates that media messages presented with certain frames impact people's perceptions about the subject of the news. As reference [20, p.52], put it, "to frame is to

select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described".

Research in the communication discipline distinguishes between frame 'building' or "how frames get established in societal discourse and how different frames compete for adoption by societal elites and journalists" [21, p.55] and frame 'setting', which deals with the effects frames exert on audiences [22]. Reference [23] argued that although the impact of frames on audience varies from an individual to an individual, nevertheless, it is rather noteworthy overall. According to reference [24, p.53], "news frames can exert a relatively substantial influence on citizens' beliefs, attitudes, and behaviours". Frame setting research predominantly focuses on an exploratory analysis of frames in media texts [6, 25] or the explanatory inquiries of the relationships between frames and audiences [23, 26].

Prior research on framing identified five predominant 'framing devices', including metaphors, exemplars, catch-phrases, depictions, and visual images that condense information and offer a 'media package' of an issue [27]. Scholars tend to examine various aspects and segments of media articles (e.g., images, title type, length of the title, news' actors, portal category, reporter angle etc.) in a search for these devices [25]. In this work, the framing theory is employed in order to conduct an exploratory analysis of media articles' titles.

III. Research Method

This research employed a qualitative content analysis of the media articles' titles. The concept of content analysis in media studies is defined as a careful observation and analysis of media interactions [28]. For the purpose of this research, content analysis was used in its traditional form – as a descriptive tool to identify main characteristics of messages conveyed through titles of the news [29].

A. Sampling Strategy

Data were collected from the two most known media outlets in the UK (i.e., BBC and The Guardian) and another two in the UAE (i.e., Khaleej Times and The National) [30, 31]. Specifically, articles' titles were examined on online portals. A typical cluster sampling approach was used to reduce a regular annual sample bias (e.g., expected reporting style and topics during holidays and expected usual yearly dynamics in reporting). Therefore, articles from 1st of June 2020 until 31st of May 2021 were reviewed and analysed. Keywords "cyber-attack" and "cybercrime" were used to search for appropriate texts (sampled *ca.* 114 texts).

B. Data Analysis

Data analysis consisted of four phases (Fig. 1). In Phase 1, the aforementioned media outlets were scanned for appropriate content (e.g., articles that focus on cybercrime victims, more specifically organisations). Seventy-two articles from UK media outlets were identified as suitable to proceed with this research (BBC n=26; The Guardian n=46), while the UAE sample consisted of forty-three articles (Khaleej Times n=30; The National n=12). Prior research demonstrated that some news headlines mislead readers with overrated or false information therefore creating incongruity between news titles and the body texts [32]. Therefore, once suitable headlines were found, it was necessary to read the texts to confirm the applicability of the actual text to the research objective.

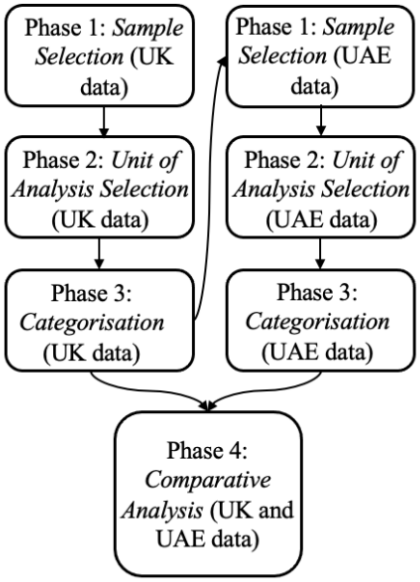


Fig. 1. Data analysis phases.

Next (Phase 2), unit of analysis was determined. For instance, a researcher could decide to analyse titles, texts, leads, quotes etc. For this paper, the unit 'title' has been singled out and analysed using categories outlined in the following step. Phase 3 involved the development of categories for a systematic analysis of titles. More specifically, the following categories were identified: (1) the title type; (2) the tone of reporting; (3) the angle of reporting; (4) the reference to victim; and (5) the connection between the title and the news content. Finally, in Phase 4 comparative analysis between the two countries was performed.

Results

Following prior research [40, 41], titles were categorised into the following types: *descriptive* (i.e., a standard category of titles that explains to readers the topic of the news); *declarative* (i.e., in addition to describing the topic, this kind of title also states what happened through the use of a specific action); *sensational* (i.e., the intention is to grab the readers' attention by causing emotions; these titles are normally shocking, bombastic and provocative); *clickbait* (these titles are similar to sensational in a sense that they use sensationalism as a style, but the major difference is that they are written with a very clear objective to increase the number of clicks even if the title is not in accordance with the text). The number of titles per a category in both samples are presented in Table 1 (in the UAE sample it was hard to distinguish between sensational and clickbait titles, therefore these categories were merged):

TABLE I. Titles' Types in UK and UAE Samples

	Descriptive	**Declarative**	**Sensational**	**Clickbait**
UK	10	51	0	11
UAE	7	26	9	

A. UK Sample

The largest part of the UK sample consisted of declarative titles. Examples of such titles are:

- "Blackbaud: Bank details and passwords at risk in giant charities hack"
- "New Zealand stock exchange disrupted by fourth 'offshore' cyber-attack"
- "Cybersecurity at risk after hackers try to sabotage Premier League transfer deal"

Forty titles out of a total of fifty-one demonstrated the negative reporter's tone and angle. Such titles therefore significantly contribute towards the building of the negative reporting frames. The negative tone and angle were mostly recognised by examining the terminology of reporting, including "risk after hackers try to sabotage" and "risk in giant hack", which gives a higher level of emotional experience to readers.

The negative aspect was also visible in the clickbait titles, for example:

- "Haunted by shame: victims of bank transfer scams tell of lasting trauma"
- "What links cybercrime, terrorism and illegal trade? Dark money"
- "Home working increases cyber-security fears"

The use of words like "haunted", "terrorism" and "lasting trauma" were employed in order to cause panic and fear among readers. These titles influenced the positioning of the topic of cyber-attacks in the domain of sensational content that relies on publicity rather than on the quality and objective reporting.

The descriptive titles were mainly related to the articles about consequences of cyber-attacks on victims and possible strategies of defence. The typical examples of descriptive titles are:

- "SolarWinds: company at the core of the Orion hack falls under scrutiny"
- "Ransomware attack on Garmin thought to be the work of 'Evil Corp'"

The majority of descriptive titles were neutral in nature although some included words like "evil", "hackers" and "danger".

The negative aspect was reflected in all titles' categories in the nature of the announcing the event itself. For example, the attacks were described as criminal activities. In addition, drastic consequences for society were explicitly stated. Statements of key actors and experts were used to support the negative element. Regarding the reference to victim organisations, 30 titles in the UK sample implicitly or explicitly pointed to the accountability and highlighted that the situation could have been prevented. An example of an explicitly negative attitude towards a victim is portrayed in the following descriptive title: "Service NSW hack could have been prevented with simple security measures". Essentially, victims of cyber-attacks are depicted as irresponsible organisations, while cybersecurity is presented as a simple task, which could not be further from the truth [7].

Most articles in the UK sample demonstrated a clear link between the title and the text; and only several titles were slightly misleading. In these instances, the primary story was used as a basis for expanding the theme by connecting it with a broader context. For example, the title "University of York: Hackers who stole data get ransom payment" gave the reader the impression that the text is about a data breach and the process of paying the ransom. Instead, the article focused on the broader issue of security breaches (e.g., how organisations handle sensitive data and the quality of post-incident communication).

Interestingly, both media outlets, BBC and The Guardian, used 'strong' words like "extorted", "victim", "violated", "suffer", "hackers" and "ruining lives" in their titles. The ultimate aim of such language is to indicate a potential threat and highlight events that resulted in the breach of sensitive data and violations of individuals' rights. In addition to 'strong' language, The Guardian journalists commonly used provocation and sometimes irony in their titles, which confirms the differences in the editorial and ownership structures between the two selected media. This is demonstrated in the following examples:

- "Poppy Gustafsson: the Darktrace tycoon in new cybersecurity era"
- "UK '95% sure' Russian hackers tried to steal coronavirus vaccine research"

B. UAE Sample

Similar to the UK sample, most of the titles in UAE media outlets were classified as declarative. Examples of such titles are:

- "Irish Department of Health target of new cyber-attack"
- "US authorities warn of 'imminent' cyber threat to hospitals"
- "UAE conducts cyberattack simulation on banking sector"

The titles did not provide specific information about the victims, except that the victims were directly listed.

As was previously mentioned, it was challenging to separate sensational and clickbait titles in this sample. Although these two categories were separated for the purpose of defining them, the empirical insight into the recorded content showed that such titles often combine both categories as evidenced in the following examples:

- "UAE: Beware! WhatsApp phishing on the rise, here's how to safeguard"
- "Coronavirus: Cyber criminals target UAE hospitals and people working from home"
- "Facebook says hackers 'scraped' data of 533 million users in 2019 leak"

It was noticeable that the aforementioned titles used popular terms like "Facebook" and "Corona" or words of warning such as "beware" and "stay alert" in order to attract the attention of readers. Reporters typically portrayed victims in a negative way by presenting them weak and unable to protect their networks. Journalists usually selected terms such as "attack", "hack", "suffer", "victim", "liability issues" and

"imminent cyber threat" to describe cyber-attacks on victim organisations.

Minority of titles in the UAE sample were characterised as descriptive. A standard example of descriptive title is:

- "Covid: Cyberattacks in UAE, GCC unlikely to subside in 2021"

Though descriptive titles are typically neutral in nature, most of the headlines in the UAE sample consisted of words such as "attack", "danger" and "threats", which contributed to the development of a negative reporting tone.

Regarding the portrayal of victims, it was typically negative. Reporters directly specified victims' identities in the titles and used words that are filled with an emotional charge (e.g., "hack", "suffer", "threat", "attack"). Positive phrases in the titles were mentioned only in the sense of protection and guidance, which further contributed to the negative framing because the victims were shown as unsafe in the matter of cybersecurity.

Articles in the sample from UAE media were generally connected to given titles, except in the case of a few clickbait headlines where journalists deliberately mislead readers, for example: "Coronavirus: Cyber criminals target UAE hospitals and people working from home." This title informs readers about a possible threat of cyber-attacks, but in the news the focus is on successful defences against cyber-attacks.

There were no significant differences in reporting practices between Khaleej Times and The National, which indicates similar editorial policies.

C. Comparative Analysis

The comparative analysis demonstrated that UK media outlets published more articles about domestic and global cyber-attacks and its victims than UAE media (UK n=72; UAE n=42). This could be due to the fact that UK outlets are more mature and experienced in reporting in general, while UAE media are still developing their practices.

Two important similarities between the two samples were observed in terms of the framing of cyber-attacks and its victims. First, the terminology of journalists was strongly oriented to the choice of words that were predominantly negative. Second, the victims were portrayed as being in eternal danger of potential cyber-attacks and yet unaware of sufficient defence mechanisms.

CONCLUSION

By analysing the titles of the sampled texts, it was possible to conclude that the media coverage of cybercrime in UK and UAE media portals was mostly shaped within a negative media frame in a given time period. Comparative analysis demonstrated similar strategies in both samples in portraying victim organisations. Reference [7] argued that negative publicity discourages victims of cybercrime (organisations, in particular) to share information about cyber-attacks, which degrades the efforts to fight cybercrime. Media could be an effective cybersecurity sharing platform, but the reporting practices must be improved.

In light of the above, future research will focus on the development of a model for an accurate cybercrime reporting, which will be distributed to the relevant regulatory bodies. Furthermore, the intention of future research is to replicate this study in various cultural environments, including Ireland. Subsequent findings will give a possibility to make a new comparation and to gain new insights about this topic. Potentially, several models for accurate media reporting could be developed due to distinct reporting approaches in various cultural environments.

ACKNOWLEDGMENT

This work was supported by Zayed University Start Up Grant [R21041].

REFERENCES

[1] J. Ramaprasad, and J.D. Kelly, "Reporting the news from the world's rooftop: A survey of Nepalese journalists", Gazette (Leiden, Netherlands), vol.65, issue 3, pp. 291-315, 2003.

[2] N. Afzal, and M. Harun, "News framing of the Arab Spring conflict from the lens of newspaper editorials," International Journal of English Linguistics, vol. 10, issue 1, pp. 352-363, 2020.

[3] L. Gui, "Media framing of fighting COVID-19 in China," Social Health Illn, vol. 43, pp. 966-970, 2021.

[4] N. Barlie, and P. Kompatsiaris, "Digital isolation and ecological abstraction. Interconnecting with environment during pandemic times", Comunicazioni Sociali, vol. 3, pp. 417-429, 2020

[5] I. Eguiagaray, B. Scholz, and C. Giorgi, "Sympathy, shame, and few solutions: News media portrayals of fetal alcohol spectrum disorders", *Midwifery*, vol. 40, pp. 49-54, 2016.

[6] M. Demata, "The language of fear: Cybercrime and the borderless realm of cyberspace in British news," I-Land Journal, vol. 1, pp. 126-144, 2017.

[7] L. Connolly, and D. Čigoja Piper, "Media framing of cybercrime: Improving victims' reporting rates", 30th European Conference on Information Systems, 2022 (accepted). Available at: https://www.researchgate.net/publication/359981133_MEDIA_FRAMING_OF_CYBERCRIME_IMPROVING_VICTIMS'_REPORTING_RATES

[8] D.I. Imamgaiazova, "Framing of cybercrimes in Russian and English media texts," *Bashkir State University* vol. 26, issue 4, pp.109-114, 2020.

[9] S. Guanah, N. Nwammuo and, I. Obi "Framing of cybercrime (Yahoo-Yahoo Business) by The Guardian and Vanguard newspapers," The Nigerian Journal of Communication, vol. 17, pp. 41-61, 2020.

[10] D. Ko, and Y. Won, "A study of effect on media exposure and cybercrime perception", Journal of Digital Convergence, vol. 14, issue 5, pp. 67-75, 2016.

[11] D. S. Wall, "Cybercrime and the culture of fear", Information, Communication & Society, vol. 11, issue 6, pp. 861-884, 2008.

[12] A. Basuchoudhary, and N. Searle, "Snatched secrets: Cybercrime and trade secrets modelling a firm's decision to report a theft of trade secrets", Computers & Security, vol. 87, 2019.

[13] F. Skopik, G. Settanni, and R. Fiedler, "A problem shared is a problem halved: A survey on the dimensions of collective cyber defense through security information sharing", Computers & Security, vol. 60, pp. 154-176, 2016.

[14] J. G. Stovall, Writing for the Mass Media, 5th ed., Boston: Allyn & Bacon, 2002.

[15] C. Condit, N. Ofulue and K. Sheedy, "Determinism and mass media portrayals of genetics. American Journal of Human Genetics," vol. 62, pp. 979-984, 1998.

[16] J. León, "The effects of headlines and summaries on news comprehension and recall," Reading and Writing: An Interdisciplinary Journal, vol. 9, pp. 85-106,1997.

[17] N. Aruguete, "Network-Activated Frames (NAF), rendering frames in a new digital era", in: Peters, M. & Heraud, R. (eds.). Encyclopedia of Educational Innovation, Springer: Singapore, pp. 1-6, 2020.

[18] A. Ardèvol-Abreu, "Framing theory in communication research in Spain: Origins, development and current situation", Revista Latina de Comunicación Social, vol. 70, pp. 423-450, 2015.

[19] E. Senocak "A framing theory-based content analysis of a Turkish newspaper's coverage of nanotechnology," Journal of Nanoparticle Research, vol. 19, issue 255, pp. 1-10, 2017.

[20] R.M. Entman, "Framing: Towards clarification of a fractured paradigm," Journal of Communication, vol. 43, issue 4, pp. 51-58, 1993.

[21] D.E. Scheufele, "Framing as a theory of media effects", Journal of Communication, vol. 49, issue 1, pp. 103-122, 1999.

[22] C.H. De Vreese, "News framing: Theory and typology", Information Design Journal & Document Design, vol. 13, issue 1, pp. 51-62, 2005.

[23] D. Tewksbury, J. Jones, M.W. Peske, A. Raymond, and W. Vig, "The interaction of news and advocate frames: Manipulating audience perceptions of a local public issue", J&MC Quarterly, vol. 77, issue 4, pp. 804-829, 2000.

[24] D. Tewksbury, and D. A. Scheufele, "News framing theory and research", in: Bryant, J. & Oliver, M.B. (eds.) Media Effects: Advances in Theory and Research, 3rd ed., Hillsdale: New Jersey, pp. 17-33, 2009.

[25] D. Čigoja Piper, Construction and reception of the autism phenomenon in print and online media in Serbia. PhD thesis, University of Belgrade, 2018.

[26] C. Schemer, "The influence of news media on stereotypic attitudes towards immigrants in a political campaign", Journal of Communication, vol. 62, pp. 739-757, 2012.

[27] W.A Gamson and L. Lasch, "The political culture of social welfare policy," in: Shimon, S. Yuchtman-Yaar, E. (eds.). Evaluating theWelfare State: Social and Political Perspectives. New York: Academic Press, pp. 397-415, 1980.

[28] K. Neuendorf, The Content Analysis Guidebook, Thousand Oaks, CA: Sage Publications, 2002.

[29] SIyengar, and A.F. Simon, "New perspectives and evidence on political communication and campaign effects", Annual Review of Psychology, vol. 51, pp. 149–169, 2000.

[30] International Media & Newspapers, "Newspaper Web Ranking 2019", 2019. Available at: https://www.4imn.com/top200/ [Accessed 1 February 2022].

[31] Statista, "Leading online news brands accessed in the United Kingdom as of February 2020" 2020. Available at: https://www.statista.com/statistics/262514/leading-online-news-brands-accessed-in-the-uk/ [Accessed 25 September 2021].

[32] S. Yoon, K. Park, J. Shin, H. Lim, H., S. Won, M. Cha, and K. Jung, "Detecting incongruity between news headline and body text via a deep hierarchical encoder", AAAI conference on artificial intelligence 2019, pp. 791-800, 2019.

[33] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[34] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[35] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[36] K. Elissa, "Title of paper if known," unpublished.

[37] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[38] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[39] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[40] J. Kuiken, A. Schuth, M. Spitters, and M. Marx "Effective Headlines of Newspaper Articles in a Digital Environment, Digital Journalism," vol. 5, issue 10, pp. 1300-1314, 2017.

[41] I. Beleslin, B.R. Njegovan, and M.S. Vukadinovic, "Clickbait titles: Risky formula for attracting readers and advertisers", 17th International Scientific Conference on Industrial Systems, 2017.